

AN INVESTIGATION OF ANSWER CHANGING ON A LARGE-SCALE
COMPUTER-BASED EDUCATIONAL ASSESSMENT

Gail Tiemann

University of Kansas

Submitted to the graduate degree program in the
Department of Psychology and Research in Education
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Chairperson: Dr. Neal Kingston

Dr. Bruce Frey

Dr. Vicki Peyton

Dr. William Skorupski

Dr. Kelli Thomas

Date Defended: April 29, 2015

The Dissertation Committee for Gail Tiemann
certifies that this is the approved version of the following dissertation:

AN INVESTIGATION OF ANSWER CHANGING ON A LARGE-SCALE
COMPUTER-BASED EDUCATIONAL ASSESSMENT

Chairperson: Dr. Neal Kingston

Date approved: May 8, 2015

ABSTRACT

Answer changing on tests has been studied for decades, however more recently answer-changing analysis has risen as an approach for exploring potential test fraud on high-stakes achievement tests. The purpose of this study was to document answer-changing patterns of students grades 3-11 on computer-based English language arts and mathematics mandated state achievement tests. Frequencies and distributions of answer-changing patterns as well as response times were reported. Additionally, relationships between student demographic characteristics and answer-changing variables were modeled using Poisson and negative binomial regression approaches. Results were consistent with prior research that has indicated large numbers of answer changes are rare occurrences that could warrant further exploration. Negative binomial regression was a better approach than Poisson regression due to overdispersion in the Poisson models. Student demographic variables were not useful in explaining answer-changing behaviors, for any of the independent variables examined. Results also add to the field's understanding of answer-changing and response-time behaviors as constructs, as well as their utility as statistical means of detecting unusual patterns on achievement tests.

ACKNOWLEDGMENTS

I would like to thank Dr. Neal Kingston for inspiring me to pursue the study of educational measurement. From you, I have learned the importance of mission-driven goals and the potential of measurement to have real impact on the educational outcomes of students. I would simply not be here without your guidance and encouragement.

I would also like to thank the other members of my committee, Drs. Frey, Peyton, Skorupski, and Thomas. Your camaraderie, input, and support have been invaluable and deeply appreciated.

My family has also been a source of motivation and encouragement. My siblings went down this path before me and made everything seem possible. My parents, Jeanine and Bob, taught me the value of education and the benefits of hard work. For these, I am grateful.

Finally, to my husband Blair Strawderman, thank you for supporting me throughout my doctoral studies. The weekends have been short and the evenings have been long, and you kept me going when I needed it the most. Thank you for sharing this accomplishment and this life with me.

TABLE OF CONTENTS

	PAGE
ABSTRACT	iii
ACKNOWLEDGMENTS	iv
LIST OF TABLES	ix
LIST OF FIGURES	xii
LIST OF EQUATIONS	xiii
LIST OF APPENDICES	xiv
CHAPTER ONE - INTRODUCTION	1
Background and Importance of the Study	1
Statement of the Problem	3
Purpose	5
Research Questions	5
Definitions of Variables	6
Significance	6
Limitations	7
Summary	7
CHAPTER TWO - REVIEW OF THE LITERATURE	9
Exploratory Research	9
Answer-Changing Analysis	10
Answer-Changing Patterns	10
Sources of Variance	13

Age	13
Gender	14
Proficiency	14
Income, Race, and Ethnicity	15
Item Difficulty	16
Other Factors Related to Answer Changing	16
Practical Reasons	17
Student Beliefs and Folk Wisdom	17
Response Time	19
Ability	19
Sub-Group Differences	19
Item-Related Factors	20
Cheating Detection	21
Statistical Models for Answer-Changing Analysis	22
Linear Models	22
Poisson Models	24
Negative Binomial Models	25
Explorations of Statistical Models and Answer-Changing Data	26
Flagging Rules	28
Summary	30
CHAPTER THREE - METHODS	31
Purpose Overview and Research Questions	31
Data Source and Participants	31

Assessment Forms and Administration	32
Response-History Logs	32
Participants	33
Procedure	35
Descriptive Analysis Procedure	35
Modeling Procedure	36
Model Selection	37
Model Comparisons	38
Item-level Analysis Procedure	39
Summary	39
CHAPTER FOUR - RESULTS	40
Answer-Changing Descriptive Statistics and Distributions	40
Total Answer-Change Frequencies	41
Wrong-to-Right Frequencies	43
Right-to-Wrong Frequencies	44
Contingency Tables	45
Answer-Changing Patterns	54
Response Time	54
Modeling Answer-Changing Variables	61
English Language Arts	62
Mathematics	66
Item-Level Analysis	70
Wrong-to-Right Changes per Item	70

Right-to-Wrong Changes per Item	71
Relationships Between Answer Changes and Item Difficulty	72
Summary	75
CHAPTER 5 - DISCUSSION	76
Answer-Changing Frequencies	76
Patterns Of Wrong-To-Right Changes	77
Response Time	77
Item-Level Answer-Changing Frequencies	78
Modeling Results	79
Additional Limitations	80
Future Research	80
Conclusion	81
REFERENCES	83
APPENDIX	92

LIST OF TABLES

	PAGE
Table 1	Number of Students Completing Summative Assessments in 2011-2012
	33
Table 2	Characteristics of Examinees as a Percentage of Total Population
	34
Table 3	Candidate Models
	37
Table 4	Statewide Student Answer-Changing Descriptive Statistics
	40
Table 5	Frequencies of Total Answer Changes – English Language Arts
	41
Table 6	Frequencies of Total Answer Changes - Mathematics
	42
Table 7	Wrong-to-Right Frequencies – English Language Arts
	43
Table 8	Wrong-to-Right Frequencies – Mathematics
	44
Table 9	Right-to-Wrong Frequencies – English Language Arts
	45
Table 10	Right-to-Wrong Frequencies – Mathematics
	45
Table 11	Changes from Wrong-to-Right by Total Item Changes – English Language Arts
	46
Table 12	Changes from Wrong-to-Right by Total Item Changes – Mathematics
	48
Table 13	Right-to-Wrong Frequencies Versus Total Item Changes – English Language Arts
	50
Table 14	Right-to-Wrong Frequencies Versus Total Item Changes – Mathematics
	51
Table 15	Right-to-Wrong Frequencies Versus Wrong-to-Right Frequencies – English Language Arts
	52
Table 16	Right-to-Wrong Frequencies Versus Wrong-to-Right Frequencies – Mathematics
	53
Table 17	Patterns of Answer Changes
	54
Table 18	Frequencies of Wrong-to-Right Item Changes 60 Seconds or Less Elapsed Screen Time – English Language Arts
	57
Table 19	Frequencies of Wrong-to-Right Item Changes 60 Seconds or Less Elapsed Screen Time – Mathematics
	59

Table 20	Wrong-to-Right Changes Per Student Changed in 10 seconds or Less Elapsed Screen Time – English Language Arts	60
Table 21	Wrong-to-Right Items Per Student Changed in 10 seconds or Less Elapsed Screen Time – Mathematics	61
Table 22	Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, English Language Arts	62
Table 23	Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, English Language Arts	63
Table 24	Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, English Language Arts	64
Table 25	Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, English Language Arts	65
Table 26	Results of AIC analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, English Language Arts	65
Table 27	Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, Mathematics	66
Table 28	Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, Mathematics	67
Table 29	Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, Mathematics	68
Table 30	Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, Mathematics	68
Table 31	Results of AIC Analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, Mathematics	69

Table 32	Number of Items per Grade, English Language Arts	70
Table 33	Number of Items per Grade, Mathematics	70

LIST OF FIGURES

		PAGE
Figure 1	Distribution of average screen time per wrong-to-right change – English language arts and mathematics	56
Figure 2	Distribution of average screen time per right-to-wrong change – English language arts and mathematics	57
Figure 3	Frequency of wrong-to-right changes per item, English language arts and mathematics	71
Figure 4	Frequency of right-to-wrong changes per item	72
Figure 5	Distribution of p-values – English language arts and mathematics	73
Figure 6	Plot of normalized of p-values versus total changes per item– English language arts and mathematics	74
Figure 7	Plot of normalized p-values versus the proportion of wrong-to-right to total changes – English language arts and mathematics	75

LIST OF EQUATIONS

	PAGE
Equation 1 Linear regression model	23
Equation 2 Poisson regression model	24
Equation 3 Akaike information criterion	38
Equation 4 Delta AIC	38
Equation 5 Akaike weight	38

LIST OF APPENDICES

	PAGE
Appendix A Model coefficients	91

AN INVESTIGATION OF ANSWER CHANGING ON A LARGE-SCALE COMPUTER-BASED EDUCATIONAL ASSESSMENT

CHAPTER ONE - INTRODUCTION¹

Evaluating student answer-changing behavior can facilitate the validity of inferences one makes from assessment scores. The purpose of this study was to explore answer-changing behaviors in the context of computer-based, high-stakes educational achievement testing. The study explored the relationship between student factors, item factors, response time, and answer changing. This chapter presents the background and importance of the study, a statement of the problem, research questions, and the significance of the study.

Background and Importance of the Study

Assessment, at its heart, represents a process for estimating what a person knows or can do (Pellegrino, Chudowsky, & Glaser, 2001). Additionally, assessment is a process of reasoning from evidence drawn from (1) models of how people understand and learn, (2) tasks that allow people to show what they know or have learned, and (3) methods for interpreting results and making valid inferences from assessment scores (Pellegrino, et al., 2001, p. 2). In the context of K-12 educational achievement testing, multiple decisions at federal, state, and local levels are made based on interpretations of examinee scores. Because of these stakes, care must be taken to establish the quality and validity of educational assessments, based on evidence, so that test results are accurate, fair, useful, interpretable, and comparable (American Educational Research

¹ Portions of Chapters One, Two, Three, and Four were presented at the Conference on Statistical Detection of Potential Test Fraud. See Tiemann & Kingston (2012, 2013).

Association, American Psychological Association, & National Council on Measurement in Education, 2014; National Council on Measurement in Education, 2012).

On educational assessments, cheating involves representing oneself or others as having knowledge via fraudulent means and in violation of rules of acceptable test taking or test administration (Cizek, 1999). When cheating occurs, assessment results are “polluted” and fairness, reliability, and validity are compromised (National Council on Measurement in Education, 2012, p. 3). Essentially, cheating destroys the interpretability and meaningfulness of test results (Cizek, 1999; Kimmel, 1997; Qualls, 2001), and denies students the opportunity to show what they know (National Council on Measurement in Education, 2012). Cheating has other consequences as well including hindered school reform efforts (Duncan, 2011), loss of public trust (Qualls, 2001), and loss of confidence in testing programs (Impara & Foster, 2006). While test developers use a number of strategies to thwart cheating *prior* to administration (Impara & Foster, 2006), opportunities to cheat *during* test administration, by students, educators, or administrators, are numerous and are serious threats to validity of test scores (Cizek, 1999; Impara, Kingsbury, Mayes, & Fitzgerald, 2005).

Professionals in education use a variety of approaches to detect cheating during test administration. While some methods simply involve human observation and proctoring of test administration, others target post-hoc, statistical analysis (National Center for Education Statistics, 2013). Many such statistical methods focus on detection of unusual patterns, or aberrance, in test data. Additionally, statistical methods evaluate the probabilities of such unusual patterns, noting whether the probability of such patterns is smaller than chance alone (Cizek, 1999). To be useful, statistical methods must be sensitive enough to detect, or flag,

anomalies, but also used cautiously enough so as to not falsely accuse examinees or test administrators (Skorupski & Wainer, 2013).

One particular example of statistical analysis involves evaluation of suspicious answer-changing patterns in test item responses. Suggested by the National Council on Measurement in Education's (2012) guidance on test data integrity as a best practice, answer-changing analysis (also known as erasure analysis) looks for higher than expected changes to answers, especially answers changed from wrong to right (Qualls, 2001). Aberrant patterns of answer changes may be indicative of changes made to items by someone other than the student themselves, such as a test proctor or educator (Qualls, 2001).

Statement of the Problem

Educators are under pressure for their students to perform well on educational assessments since the consequences of those scores can be critical (Amrein-Beardsley, Berliner, & Rideau, 2010; Qualls, 2001). Scores on mandated state tests can be attached to graduation, admission to college, as well as educator, school, or district performance evaluation, teacher salaries, and school funding (Cizek, 1999; Kimmel, 1997). Because of this intense pressure to perform, educators may engage in cheating behaviors that are not typical of them (Rideau, 2009). Some researchers have argued that cheating among educators has increased since the implementation of No Child Left Behind, though how much no one is certain (Amrein-Beardsley et al., 2010).

However, recent cases of cheating have certainly caught the attention of newspaper headlines, most notably an extensive cheating scandal in the Atlanta Public Schools. In this situation, statistical detection of cheating via erasure analysis (among others) was implemented and then followed up by a thorough investigative process, where ultimately 178 teachers and

principals confessed to changing student answers at 44 schools (Severson, 2011). Erasure analysis has been used in other states as well. The New York State education department conducted erasure analysis on the high school Regents exams. In this example, one of 64 flagged incidents resulted in the termination of a Bronx assistant principal (Otterman, 2011). This particular analysis showed that of 1,013 items erased on the exam, 94% had been changed from wrong to right, where typically in that testing program, about 50% of answers were changed from wrong to right. Related to the application of erasure analysis in testing programs nationally, a USA Today survey (Bello & Toppo, 2011) reported that 20 states and Washington, D.C. routinely conduct erasure analysis on all student tests.

In response to current events around test score integrity, U.S. Secretary of Education Arne Duncan (2011) in letter to Chief State School Officers, urged administrators to make assessment security a high priority, noting that security and data quality were “essential elements of an assessment system” (para. 3). In response to and support of this call, The National Council on Measurement in Education (NCME) released guidelines on testing and data integrity, in efforts to steer state education agencies (SEAs), assessment consortia, test publishers, and contractors toward recommended practices, policies, and procedures (National Council on Measurement in Education, 2012). Along with these best practices, NCME additionally recommended consistent evolution of our test and data integrity methods, noting the need for improved real-time and post-hoc statistical anomaly detection techniques (p. 6).

The importance of the validity and integrity of high stakes achievement test scores calls for the professionals in education and the research community to pay attention to the problem of cheating (Cizek, 1999; Qualls, 2001) and to continue evolving cheating detection methods. While answer-changing analysis has been used as a statistical method for detecting cheating

behavior in schools, the practice has received little attention in the research literature (Bishop, Liassou, Bulut, Dong, & Stearns, 2011). Additionally, literature that does exist has focused solely on paper-pencil tests, using either hand or optical scanner detection of changed item responses. Given the nature and consequences of high-stakes assessments and cheating detection, interest and attention to the topic is important and warranted (Qualls, 2001). Additionally, the growth of computer-based testing in the K-12 arena provides additional context for study since more information on student-response patterns exists than has ever been available before. Indeed, ample opportunity exists for investigation of answer changing as a construct and delving deeper into its use as a method for flagging suspicious item-response patterns.

Purpose

The purpose of this study was to document answer-changing patterns of students grades 3-11 on computer-based English language and mathematics arts mandated state achievement tests. Results add to the field's understanding of answer-changing and response-time behaviors as constructs as well as their utility as statistical means of detecting unusual patterns on achievement tests.

Research Questions

The study addressed the following questions:

1. How can frequent answer-changing on computer-based achievement tests be understood?
2. How do student and item difficulty relate to answer-changing on computer-based achievement tests?
3. What statistical models and distributions are appropriate for answer-changing variables and analysis?

Definitions of Variables

Answer change

An answer change was recorded when a test taker changed from one answer choice to another.

Wrong to Right

While a student could change a response multiple times, a wrong-to-right answer change was noted when the very last change in the student's pattern changed from a wrong answer to a right answer, regardless of how the student initially answered the question.

Flagging Rule

A flagging rule represented a threshold value at which point an unusual or unlikely response pattern was suspected.

Response Time

Response time reflected the elapsed time that a student spent marking an answer choice.

Significance

While answer changing has been studied in the context of paper-pencil tests for decades, there is a paucity of published research related to answer changing in the context of large-scale state accountability assessment. Moreover, previous research has been limited due to reliance on scanner-based erasure detection methods; answer-changing cannot actually be seen on paper-pencil tests, only inferred from detection of light marks on paper. Additionally, very little has been documented about the construct related to computer-based tests. Given that computer-based assessment provides data for richer analysis of answer changing than has ever been possible before, this study expanded the field's knowledge of this construct.

Since state education agencies (SEAs) are primarily responsible for monitoring test score integrity (National Center for Education Statistics, 2013; National Council on Measurement in Education, 2012), states must be prepared with information and methods that are practical and useful in detecting aberrant item-response patterns *before* questionable situations arise. To this end, developing a thorough understanding of expected as well as unlikely patterns of answer changing is critical. This study uniquely responds to this critical need by documenting answer-changing patterns and building theory about this evolving construct.

Limitations

Aberrance flagged using statistical methods does not represent truth; information from such analyses are limited and are essentially “incapable of detecting anything other than some occurrence that is out of the ordinary or different from other occurrences” (Cizek, 1999, p. 136). In fact, unusual patterns could be caused by chance or by students using legitimate test-taking strategies (National Center for Education Statistics, 2013). For example, entire classes of students might be explicitly taught by educators to mark through all unlikely answers before selecting a final answer (Wainer, 2012). Such a practice could be mistakenly associated with cheating. Thus, statistical detection then should serve as one component in a comprehensive evaluation approach by those embarking on investigation of potential cheating or test fraud (National Center for Education Statistics, 2013).

Summary

This chapter presented the background of the study, a statement of the general research problem, the study’s purpose, research questions, significance, and potential limitations. Chapter Two describes the literature relevant to the topic, including results from various answer-changing analysis studies, sources of variance related to answer-changing, a review of response time

studies, sources of variance related to response time, and statistical models for answer-changing analysis. Chapter Three presents the exploratory research design including data sources, variables, participant demographics, and a summary of the data analysis procedures. Chapter Four presents the study's results. Chapter Five discusses the results in the context of previous literature, describes study limitations, possibilities for future research, and overall conclusions.

CHAPTER TWO - REVIEW OF THE LITERATURE

This chapter provides the reader with an introduction to exploratory research as well as results of relevant answer-changing studies. This is followed by a description of sources of variance related to answer-changing that have been noted in the literature. The second section describes results from response-time studies including sources of variance related to response time and how response time has been utilized in cheating detection. The third section describes statistical models that might be used in answer-changing analysis and how use of these models may affect cheating detection results.

Exploratory Research

This study primarily utilized a quantitative-exploratory approach. Quantitative-exploratory research uses inductive approaches to derive deep understanding of the topic under study (Stebbens, 2001). With this approach, inductive approaches and reasoning move from the particular to the general, deriving generalizations through weighing evidence, judging plausibility, or arriving at conclusions or beliefs (Nickerson, 2010). Theory then, is derived from the data and serves as catalyst for new research and further exploration (Stebbins, 2001). Essentially, new theory leads to new strategies and methods of exploration with their “own rules yet to be explored” (Glaser & Strauss, 1967, p. 187). This paradigm of exploration and derived generalizations is the foundation of grounded theory methodology.

While one might think of grounded theory as purely a qualitative approach, the concept equally applies (although less often) to exploratory quantitative methods as well (Stebbens, 2001; Glaser & Strauss, 1967). In fact, Glaser and Strauss stated in their seminal work on grounded theory that quantitative data could be a very rich medium for discovering theory. With quantitative approaches, theory can be derived from looking for general relationships between

concepts, either positive or negative. Furthermore, a theory-generating approach need not commence with preconceived hypotheses. Instead, analysis leads to posing new hypothesis to be evaluated in future work (Milliken, 2010).

While broad-ranging, exploratory research is not simply a “fishing expedition.” Instead, exploratory approaches are structured and follow guidelines in order to be purposive and systematic, yet also flexible and open-minded (Stebbins, 2001). Exploratory research then, is a scientific process (Vogt, 1999). Stebbins (2001) likened the process to setting and implementing a meeting agenda with points specified in advance, yet flexible enough to adapt to discussion and new ideas not previously declared.

Answer-Changing Analysis

Answer-changing analysis involves examining answer-changing behavior on multiple choice tests and looking for larger than expected counts or unusual patterns of basic answer changes or wrong-to-right answer changes (National Center for Education Statistics, 2013). Aberrant patterns of answer changes may be indicative of changes made to items by someone other than the student themselves (Qualls, 2001).

Answer-Changing Patterns

Although limited, answer changing has been documented in the research literature in context of statistical cheating detection. Qualls’ (2001) study described baseline answer-changing and wrong-to-right frequencies for large-scale, K-12 achievement tests in both low- and moderate-stakes assessments. Qualls’ first study examined low-stakes, paper-pencil results from 16 districts in Iowa during the 1994-1995 school year. Results showed that “more than 90% of students changed three or fewer responses” (p. 12), with about 50% changing one response. Across the content areas and grades represented, 50% of students erased at least one answer with

38-64% erasing zero answers, 20-29% erasing one answer, and 7-16% erasing two answers. In addition, when students changed only one answer, 50% were wrong to right and about 20% were right to wrong. In general, as the number of answer changes increased, the number of wrong-to-right changes decreased.

In Qualls' second study, results from moderate-stakes assessments were compared to results from low-stakes assessments. Qualls found that answer-changing patterns were similar between the two assessment types. Qualls concluded that "it would be rare to see a student change more than 15% of the items" (p. 15) and that "wrong-to-right changes would not typically exceed 55% for one erasure, and for multiple erasures the number of 100% wrong-to-right changes would be dramatically lower" (p. 15). Ultimately, Qualls felt that answer-change counts and wrong-to-right change counts above these thresholds could be used to flag aberrant tests.

Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) also examined answer changing in the context of large-scale, K-12 achievement testing. Their study examined responses from four state, paper-pencil testing programs and reported the proportions of total items that were changed in general, as well as the proportions that were changed from wrong to right. Total proportions of answers changed ranged from .002 to .166. Wrong-to-right change proportions ranged from .001 to .116. Overall, in this data, answer changes occurred about 2% of the time and wrong-to-right changes occurred about 1% of the time.

Both the Qualls (2001) and the Primoli et al. (2011) studies were based on paper-pencil tests with answer changes detected via optical scanning equipment. Primoli et al. cautioned readers that answer-changing counts could vary by program, since optical scanning sensitivity

settings often vary. Thus, the potential errors in detecting answer-changing counts is a weakness of paper-pencil tests and optical scanning in general.

While research related to answer changing in the context of cheating detection is limited, answer changing in the context of general paper-pencil tests has been studied for decades. For instance, numerous researchers have noted basic answer-changing frequencies. Benjamin, Cavell, and Shallenberger (1984) synthesized the research in this area finding that across 15 studies, the percentage of test takers changing one or more answers ranged from 57% to 96% with a median of 84%. Benjamin et al. also noted about 16% of students changed no answers at all and that across 18 studies, answer-changing frequencies were “very consistent” (p. 136). Later studies were consistent with Benjamin et al.’s synthesis. McMorris’ (1991) found that 75% of test takers changed at least one answer and Geiger (1991) found 97%.

Researchers have also noted the average percentage of items changed per student. Again, Benjamin et al.’s (1984) research synthesis found that across 28 studies, the proportion of items changed was small, ranging from 2.2% to 9.0% with a median of 3.3%. A few later studies showed results consistent with Benjamin et al. McMorris and Weideman (1986) found that 5.28% of items were changed across items and people. Prinsell, Ramsey, and Ramsey (1994) found a change rate of 4% and Geiger (1991) found 4.4%. In Al-Hamly and Coombe’s more recent study (2005), the average number of items change per student was 2.65%.

Several researchers also noted wide variances for average answers changed per student. Geiger found large standard deviations for number of changes with some being “greater than the mean” (1991, p. 253). Jackson (1978) studied three samples of youth age 10 – 14 years, finding average changes per item of 10.36 (*SD* 5.53), 10.57 (*SD* 5.21), and 10.02 (*SD* 6.56). McMorris and Weideman (1986) noted a mean percentage of items changed of 6.9 (*SD* 4.2).

Sources of Variance

In addition to documenting basic answer-changing frequencies and patterns, researchers have studied relationships between answer changing and a variety of other student- and test-related factors.

Age

Answer-changing research has addressed a variety of student populations and age groups. A number of studies sampled students in higher education settings and targeted answer changing by test takers. Early small-scale studies involved undergraduate nursing, business, accounting, and statistics students with sample sizes ranging from 50 to 300 students (Geiger, 1991, 1997; Green, 1981; Jacobs, 1972; McMorris & Leonard, 1976; Nieswiadomy, Arnold, & Garza, 2001; Prinsell, et al., 1994). Other studies involved graduate psychology and education students (McMorris & Weideman, 1986). Only a few studies have focused on answer changing by children. Casteel (1991) reported research by Crocker and Bensen (1977) which found that seventh-grade students made fewer answer changes than older examinees. McMorris et al. (1991) studied answer changing in rural fifth- and sixth-grade students, finding that proportions of wrong-to-right, wrong-to-wrong, and right-to-wrong changes were similar in proportions to those found in adults. In a more recent, large-scale study in the K-12 achievement testing context, Primoli et al. (2011) found similar answer-changing patterns across students in grades 3–11. While several studies have examined answer changing across age groups, no distinct patterns seem to have emerged from this work. Additionally, most studies were with older students in higher education environments.

Gender

Research on answer changing has also considered the effect of gender. Reile and Briggs (1952) found that females change their answers more often than males. Skinner (1983) found a significant difference between females and males, with females changing answers twice as frequently. McMorris et al.(1991) also found that a “somewhat high” proportion of females changed answers more often than men (82% vs. 69%) (p. 11). However, Al-Hamly and Coombe (2005) found that males made more answer changes (2.86 to 2.45 changes per person) though the difference was not statistically significant. Mueller and Shwedel (1975) also found that males made more answer changes. While some research reported an effect that favored either females or males, Geiger (1991) found no significant difference at all between genders. In terms of wrong-to-right or right-to-wrong changes, Al-Hamly and Coombe (2005) found that females made more wrong-to-right changes. However, Mueller and Shwedel (1975) found that males made more right-to-wrong changes. Across many years of studies, results are mixed. Additionally, all of the studies mentioned examined answer-changing frequencies on paper-based tests only.

Proficiency

Studies have also examined the relationship between student proficiency and answer changing, with proficiency defined in a variety of ways including test total score, course grades, external standardized measures, and self-reported academic performance. Results have been mixed. Benjamin et al.’s (1984) review cited six studies with statistically significant, negative relationships between test scores and the number of answers changed. Five other studies in the review reported “nonsignificant results in the same direction” (Benjamin et. al., 1984, p. 137). Other studies have also reported mixed results. Best (1979) found that students with higher

grades tended to change answers less frequently than students with lower grades, and the former made fewer right-to-wrong changes than other students. Al-Hamly and Coombe (2005) also found that higher scorers changed answers less frequently than other students; however, McMorris et al. (1991) did not. Again, all of the tests administered in these studies were paper-pencil based and administered to a variety of audiences, mostly in higher-education contexts.

Primoli et al. (2011) examined the relationship between answer-changing and ability, with ability represented by a testing program's item response theory-based ability metrics. In these results, abilities ranged from -2.357 to 6.717 (p. 17) and showed strong, cubic relationships between ability and total erasure proportions, as well as between ability and wrong-to-right erasures. Thus, total answer-changing proportions increased as abilities increased, up to a point. Past the point, the total answer-changing proportion decreased as ability increased. While the authors cautioned against over-generalizing results by grade, patterns were similar across grades and programs.

Income, Race, and Ethnicity

Income, race, and ethnicity are not well-represented in the literature as potential sources of variance in answer-changing analysis. Matter (1986) studied answer-changing patterns of elementary students on the Iowa Test of Basic Skills, recording right-to-wrong, wrong-to-right, and wrong-to-wrong frequencies by ethnic group, family income, and achievement level. While second graders and African American students had the highest mean answer changes, Matter found no significant difference between low and non-low income students. In their analysis, Primoli et al. (2011) plotted average wrong-to-right proportions against school-level percentages of economically disadvantaged students, finding a quadratic relationship across grade levels and content areas. The authors also examined variance in answer-changing rates across ethnic groups

finding that African American and Hispanic students had higher rates of answer changing than students of similar abilities in other ethnic groups. The results were “especially true for middle to high ability students. This was true for both Reading and Mathematics” (p. 22-23).

Item Difficulty

Researchers have examined the relationship of item difficulty to answer-changing behaviors. Item difficulty has been defined most often as p-values from sample data or from data on previous test administrations. Other researchers defined difficulty based on position above or below the median percent correct for all items or through whole test judgment ratings (Benjamin et al., 1984). Jacobs’ early study (1972) found that the easiest items were changed the least frequently. These results were confirmed by Green (1981) and McMorris and Weideman (1986). Green’s study also found that difficult items were changed more frequently, regardless of students’ levels of test anxiety.

Wrong-to-right answer changes have also been examined. Beck (1978) found that easier items were significantly more likely to be changed from wrong to right than hard items. Green (1981) also found a significant main effect for item difficulty group (difficult, moderate, or easy) with both number of wrong-to-right changes and total number of changes. Al-Hamly and Coombe (2005), more recently, found a significant positive correlation of .217 between item difficulty and wrong-to-right changes (p. 518).

Other Factors Related to Answer Changing

In addition to formal study of demographic variables and their relationships to answer changing, researchers have considered other, more general factors in the literature as well.

Practical Reasons

One researcher explored why students change answers on tests in general. As part of a study examining the impact of instruction about the benefits of answer changing on answer-changing frequencies, McMorris and Weideman (1986) surveyed students, asking them to choose among five possible reasons for changing answers. In this sample, 57% changed answers because of rethinking the item and coming up with a new answer. Twenty-eight percent reread the item and developed a better understanding of the question. Other reasons included clerical error (8%), finding a clue (3%), and learning from an item appearing later on the test (3%) (p. 96).

Student Beliefs and Folk Wisdom

Several authors have explored student beliefs about changing or not changing answers on multiple choice tests. Mathews reported as early as 1929, that a majority of students felt they would *not* benefit from changing answers (Mathews, 1929). Benjamin, et al.'s 1984 review of the literature (1984) supported this notion finding that across many years of research results, "approximately three out of every four of these students felt answer changes would lower their scores" (p. 133). Additionally, while perceptions may exist related to changing answers, Geiger (1991) reported that students were actually poor predictors of the impact of their own answer changing. In his study of 127 undergraduate accounting majors, only 26% correctly predicted the outcome of changing their answers, with 11% overestimating and 65% underestimating their results.

Nieswiadomy et al. (2001) expressed that the folk wisdom related to staying with "first impressions" (p. 142) has probably been "handed down by word of mouth and through written instructions that frequently accompany examinations" (P. 142). Benjamin et al. (1984) and

Prinsell et al. (1994) added that peers and teachers are a likely source for the “admonition on answer changing” (Benjamin, et al., 1984, p. 133). Mathews’ early study (1929) also noted examples of teachers directly telling students not to change their answers. Prinsell et al. noted recently as 1994, that students were indicating that they had been told not to change answers on tests (p. 328).

Other authors have explored beliefs about answer changing more thoroughly. Specifically, if students were taught explicitly that changing an answer tends to result in score gains, would beliefs about answer changing and scores also change? McMorris’ research in the mid-80’s explored these factors. McMorris and Weideman (1986) considered that if test takers were generally reluctant to change answers, they might only change answers when they were most confident, thus explaining the prevalence of test score gains from answer changing. On the other hand, if an understanding of the empirical benefits of answer changing were conveyed to students, perhaps more answer changes would occur and the prevalence of gains would be erased.

In a 1986 study, McMorris and Weideman included information about previous empirical results as part of a graduate course curriculum. After this instruction, in general, students still gained from changing their answers. Also, the frequency of answer changing between samples of students who were instructed and who were not instructed on the benefits of answer changing was “essentially equivalent” (p. 135).

Prinsell et al. (1994) conducted a similar study with 300 undergraduate and graduate students. Students were surveyed before and after instruction about the benefits of answer changing. Results indicated that while student attitudes toward answer changing were more favorable after instruction, there was no corresponding increase in scores. Prinsell et al.’s and McMorris and Weideman’s results align with an earlier conclusions stated by Mueller and

Wasser (1977), namely that “there appears to be no systematic relationship between percent of answers changed and the nature of directions given to students” (p. 10).

Response Time

The amount of time test takers use to answer an item has been studied in psychological research for many years (Schnipke & Scrams, 2002). A number of factors can influence response time including examinee, item, and contextual factors (Wise & Kingsbury, 2006).

Ability

Schnipke and Scrams’ (2002) review of the literature in this area presented a variety of empirical results. Related to examinee factors, Swanson, Featherman, Case, Luecht, and Nungester (1999, as cited in Schnipke & Scrams, 2002) found that lower-ability examinees spent about the same amount of time answering items at all levels of difficulty. In contrast, higher-ability examinees typically spent more time on harder items and less time on easier items. However, Swanson et al.’s research did not find a statistically significant relationship between mean response time and ability. Using data from the computer-adaptive Graduate Record Examination (GRE), Schaeffer (1995) also found varying patterns in response time among students in different ability categories. Results indicated that across three sub-tests, lower-ability examinees spent about the same amount of time on the easiest and the most difficult items. Middle-ability examinees spent more time on harder items. High-ability examinees spent “considerably-more” time on difficult items (Schaeffer, 1995, as cited in Schnipke & Scrams, 2002, p. 256).

Sub-Group Differences

Other researchers have studied sub-group differences in response times, with most comparing median or mean response times (Schnipke & Scrams, 2002) and one team utilizing

survival analysis methodology (Schnipke & Pashley, 1997). Schnipke and Scrams (2002) report in their research synthesis that results related to differences in gender and ethnicity are mixed. Some studies have reported that gender and ethnicity are not significant predictors of response time and other results have found “small differences” (2002, p. 259). Schnipke’s (1995) results also showed that rapid-guessing behavior varied by gender and test type; males demonstrated more rapid guessing on an analytical test and females more on a quantitative test. Schnipke found no gender difference on a verbal test.

Item-Related Factors

Related to item factors, several researchers have explored the relationship between item difficulty, item discrimination, item length, and response time. Halkitis’ (1996) study reported that item length (number of words), item difficulty, and item discrimination (point-biserial correlation) explained 50.18% of the variation in log response times on a computer-administered licensing exam. Smith (2000) examined similar variables using responses from the Graduate Management Admissions Test (GMAT), a computer-adaptive exam. Using multiple regression, Smith found that between 35.4% and 71.4% of the variability in response time across item types could be explained by item word count, item difficulty, and item discrimination. For two GMAT item types, the relationship between item difficulty and response time was statistically significant and quadratic. For two item types, one relationship between item discrimination and response time was statistically significant and quadratic and the other was cubic. Word count demonstrated a statistically significant linear relationship for all item types except one, reading comprehension. Smith’s research underscored that item type should be considered, in addition to other variables, in response time analysis.

Cheating Detection

Wise and Kingsbury (2006) used response time to detect compromised items on the National Council Licensure Examination (NCLEX). Their assumption was that shorter item-response times would result from examinees with advance knowledge of an item's content and correct answer. Simply recalling the correct answer would "require substantially less time than needed to read the item, understand its challenge, and do the mental processing needed to identify the correct answer" (p. 2). Wise and Kingsbury's analysis compared items administered during pilot testing, which could not be affected by exposure, to later test administrations. For items that registered as "too easy" and thus possibly compromised, the authors plotted the average response time for each item, for both correct and incorrect responses. Results indicated that correct responses took less time than incorrect responses for "too easy" and thus potentially compromised items. The average response time difference for "too easy" items was 12.66 seconds (*SD* 8.23) (p. 10). The average response time difference for other items (not suspected of being compromised) was 9.47 seconds (*SD* 9.10). The differences were not statistically significant ($p < .10$), but the potentially compromised items had a larger correct to incorrect time difference in 27 out of 31 comparisons (p. 10).

Using a mixture model of two lognormal distributions, Schnipke and Scrams (1997) explored differences in examinee behaviors for "solving" items in contrast to "rapid guessing" on non-adaptive computer-based tests. The lognormal distribution has been indicated for use with item-response times since it provides the best fit (Schnipke & Scrams, 1999) and since response-time distributions "tend to be skewed" (Smith, 2000). In the Schnipke and Scrams study (1997), "solution behavior" was indicated when examinees actively sought to answer an item correctly. "Rapid guessing behavior" was indicted when examinees answered questions

rapidly, as time for the test ended. Results of their modeling research indicated that the rapid-guessing distribution was similar across items, suggesting that the behavior is unaffected by item content. Additionally, Schnipke and Scrams (2002) stated that the mixture model could be applied to other application of test-taking strategy, including the concept that “stolen’ items might have fast, correct responses” (p. 249).

Finally, the National Center of Education Statistics (2013) noted that response time tracking on computer-based tests provides another source of information for potential cheating detection. NCES noted that short and long response-time patterns on difficult versus easy items could be used to detect item exposure or item coaching. Specifically, response-time tracking on computer-based tests provides additional information simply not retrievable from paper-pencil test administrations.

Statistical Models for Answer-Changing Analysis

Answer changing generates count data. Whether recorded as a simple yes or no, wrong to right, or right to wrong, etc., each instance of an answer change is counted discretely and thus can only take the value of zero or a positive integer. Count data is very common in psychological studies with many researchers seeking to explain variation in the number of event occurrences using predictors (Coxe, West, & Aiken, 2009; Gardner, Mulvey, & Shaw, 1995).

Linear Models

When modeling count data, researchers may be tempted to use ordinary least squares (OLS) regression. However, using OLS regression with count variables as outcomes causes several problems. Unless the mean of the outcome variable is above 10 (a rule of thumb), using OLS regression may produce biased standard errors and significance tests (Coxe, et al., 2009) as

well as “nonsensical, negative predicted values” (Gardner, et al., 1995, p. 393). These problems stem from violations of OLS regression assumptions.

Ordinary least squares regression is used to predict values of dependent variables, also known as criterion variables, based on their association with values of independent variables, also known as predictors. In the case of multiple linear regression, an equation utilizing a linear combination of predictor variables is sought that also minimizes errors of prediction. The form of multiple OLS regression with two predictor variables is

$$\hat{y} = b_1x_1 + b_2x_2 + a \quad (1)$$

where \hat{y} is the predicted value for y , x_1 and x_2 are the predictor variables, b_1 and b_2 are the regression coefficients, and a is the intercept, or the value for y when x_1 and x_2 are both zero (Segrin, 2010; Tabachnick & Fidell, 2007).

Assumptions must be met in order to use this form of regression. First, the data points under analysis should be independent; the data from one participant should be independent from the data provided by another participant (Coxe, et al., 2009; Segrin, 2010). Next, for each observed value of a predictor variable, the corresponding values of the criterion variable should be normally distributed (conditional normality). Finally, for each value of a predictor, the variance of the distribution of prediction errors ($\hat{y} - y$) must be the same (homoscedasticity) (Coxe, et al., 2009; Gardner, et al., 1995; Segrin, 2010).

Count variables often violate the principles of homoscedasticity and conditional normality. For example, a count variable may have larger conditional variance as the value for the predictor becomes larger. This is often a result of lower counts, and thus higher variability, at higher values of the predictor (Coxe, et al., 2009). Heteroscedasticity, then is a problem with model fit and leads to “biased standard errors and biased tests of significance” (Coxe, et al., 2009,

p. 122). Additionally, the distributions of count variables often violate assumptions of conditional normality, since they are often positively skewed; count variables have no values less than zero and many “low-count observations” (p. 122). Violations of these two assumptions make values of regression coefficients and standard errors, as well as results of significance tests more difficult to interpret.

Poisson Models

Poisson regression belongs to the family of generalized linear models and can be a natural fit for count data and rare events (Cameron & Trivedi, 1998; Cox, et al., 2009; Gardner, et al., 1995; Kato & Bart, 2010). The form of the Poisson regression model is

$$\ln(\hat{\mu}) = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p \quad (2)$$

where $\hat{\mu}$ is the predicted count of y given $x_1 \dots x_p$ values of the predictors, \ln is the natural logarithm, b_0 is the intercept, and b_p etc., are the regression coefficients (Cox, et al., 2009, p. 123). Note that the criterion in the above equation is not a count, but a natural logarithm of a count. Also, the distribution of y given $x_1 \dots x_p$ follows the Poisson distribution.

The Poisson distribution has properties which make it a better fit for count data than the normal distribution. First, it is a discrete distribution which provides probability values for only nonnegative integers. Thus conceptually it is aligned with count variables which also must be nonnegative. The Poisson distribution is also specified by a single parameter μ which represents the mean number of occurrences as well as the variance; thus the mean and the variance of this distribution must be equal. From this property, it follows that as the distribution mean increases, the variance also increases. Increasing variances with increasing means is a common characteristic of count data. Finally, for Poisson regression, the distribution of prediction errors ($\hat{y} - y$) follows the Poisson distribution. Since under the Poisson distribution variances increase

with larger expected values of x (means), the variances of the errors need not be the same across all values of predictors and can be heteroscedastic, which is not allowed by ordinary linear regression models.

There are situations where Poisson regression may not adequately fit count data. As mentioned above, the Poisson model requires that the mean and variance are equal. However, count data may exhibit more variability than is allowed in the model. When the data exhibit a variance which is larger than mean, the model is *overdispersed* (Coxe, et al., 2009; Gardner, et al., 1995). Overdispersion can result from counts which are not independent from each other. If overdispersion is not accounted for, standard error values will be too small and parameter estimates will be too large (Coxe, et al., 2009, p. 130; Tang, He, & Tu, 2012).

Negative Binomial Models

The negative binomial regression model, another model in the family of generalized linear models, is another approach for modeling count data (Coxe, et al., 2009; Gardner, et al., 1995; Tang, et al., 2012). While the Poisson regression may provide better model fit for count data than OLS regression, some problems may still exist. For instance, the Poisson regression model does not allow for an error term, as is provided by OLS Regression. Thus, “the Poisson model does not allow heterogeneity among individuals” (Coxe, et al., 2009, p. 132) above and beyond what is explained by the predictors. As a result, there may exist larger variances in the model than would be expected in the Poisson distribution.

In contrast, the negative binomial model allows more flexibility to consider individual variances. Rather than assuming that all individual counts come from the same Poisson distribution, the negative binomial model allows individuals with the same values on the predictor variable to be modeled from Poisson distributions with different means (Coxe, et al.,

2009). The variation in the means among the Poisson distributions follows the gamma probability distribution (Coxe, et al., 2009; Tang, et al., 2012). The predicted mean of the negative binomial model, given the predictors will be the same as that produced by the Poisson distribution, however the variance will be larger.

Where the Poisson model had a single parameter, the negative binomial model has two, a scale parameter μ and a dispersion parameter α (Coxe, et al., 2009; Tang, et al., 2012). The mean of the negative binomial model is equal to μ and the variance is equal to $\mu(1 + \alpha\mu)$. The α parameter represents the dispersion (similar to ϕ in the overdispersed Poisson regression model). If the value for α is close to zero, the negative binomial distribution is very close to the Poisson distribution. If $\alpha > 0$ then the distribution is overdispersed. Additionally, as the value for α gets larger, there is more variability in the data than what can be explained simply by the Poisson distribution (Tang, et al., 2012). The negative binomial model and its parameters can be estimated using maximum likelihood estimation with commonly available statistical software packages.

Explorations of Statistical Models and Answer-Changing Data

Bishop, Liassou, Bulut, Seo, and Bishop (2011) explored model fit related to several different answer-changing analysis procedures. At each grade level evaluated, the authors found that for basic wrong-to-right counts, distributions were extreme and positive, thus ruling out the normal distribution. Additionally, distribution shapes were similar at all grade levels.

Related to basic answer-changing counts, Bishop et al. noted that two characteristics of this data may interfere with the fit of the Poisson distribution. First, answer-changing counts are often zero inflated, or have a higher frequency of actual zero values than the Poisson model would predict. Also means and variances are often not equal as required by the Poisson model.

Because of these factors, the authors compared the fit of the Poisson model to the negative binomial model, finding that the negative binomial distribution fit their data better. However, the authors did not report fit of corrected Poisson models such as the overdispersed and zero-inflated Poisson models.

Bishop et al. also examined the distribution of wrong-to-right to total-change ratios, since this ratio can be used as a dependent variable in answer-changing analysis. The authors noted that for many students, total change count was zero which can produce a ratio of 0:0, an undefined value in mathematics. How this outcome is treated affects the resulting descriptive statistics and distribution of ratios. When Bishop et al. treated the zero denominator as missing, means and variances were “similar across grade levels” (p. 17). In most grades, the same was true if the ratio was simply recoded to zero. However, distributions varied depending on the divide by zero procedure used. When Bishop et al. treated zero denominators as missing, distributions were *negatively* skewed with most means equal to about 0.6. Ratios that were recoded to zero resulted in *positively* skewed distributions, with means equal to about 0.3.

Noting that conditional relationships between total changes and wrong-to-right changes could be used as a flagging rule, Bishop et al. also examined the fit of linear and Poisson regression models to the conditional relationship of the two. For the linear regression model, the authors reported that total changes explained a significant proportion of variance in wrong-to-right changes across all grades. However, a plot of the residuals demonstrated heteroscedasticity (non-constant differences between the actual values of wrong-to-right changes and the predicted values of wrong-to-right changes); the variance in the residuals tended to increase as total changes increased. Because of this property of the data, the authors also fit two Poisson regression models, a null model (a model containing a constant for wrong-to-right counts across

all values of total changes) and a model adding in total changes as a predictor. They found that a Poisson model containing total changes fit the data better than a null model alone and concluded that the Poisson regression model “good fit” at the student level (p. 50).

Flagging Rules

Answer-changing analysis has typically involved use a flagging rule, where patterns of changes that exceed a pre-defined threshold are marked for further review. The flagging rules are generally based on a comparison of actual versus expected student answer-changing behaviors.

In practice, state education agencies and assessment contractors have used various flagging rules in applied answer-changing analysis. In one analysis conducted in a district in Morgan County Georgia, students were flagged if the count of wrong-to-right answer changes exceeded the state average plus three standard deviations (Schiliro, 2010). Other flagging rules are more conservative. Louisiana Test Security policy (Erasure Analysis, 2012) offers a flagging rule of the state average of wrong-to-right answer changes plus four standard deviations. Recently, researchers at the Data Recognition Corporation applied a flagging rule based on the state average of wrong-to-right changes plus eight standard deviations (Primoli et al., 2011).

Bishop, Liassou, Bulut, Dong, and Stearns (2011) compared several student-level flagging rules, noting the number of students flagged by each. Basic wrong-to-right counts, wrong-to-right to total-change ratios, as well as linear and Poisson regression-based rules were compared.

First, Bishop et al. compared simple count-based flagging rules. For example, if a researcher used a flagging rule based on 0.05 significance level for basic wrong-to-right counts, students with 2 or greater wrong-to-right answer changes would be flagged under the normal distribution and 3 or greater would be flagged under the Poisson and negative binomial

distributions. Bishop et al. found in their actual data though, that a cut point of 2 (normal distribution) would flag 7% of students rather than the expected 5%. At a cut point of 3, the Poisson distribution would predict that less than 1% of students would be flagged and the negative binomial distribution would predict about 3% of students would be flagged. In the actual data though, about 2.8% of students were flagged at a cut point of 3. Thus the negative binomial model was the closest to actual frequencies.

Next, Bishop et al. plotted wrong-to-right counts against wrong-to-right to total-change ratios, flagging the most extreme 2.5% of cases. However, the authors note that “even the lowest [wrong-to-right] counts can have high [wrong-to-right to total change] ratios” and that “small [wrong-to-right to total-change] ratios (about 0.30) can have large [wrong-to-right] counts” (p. 6). Thus the authors urge care in using and interpreting wrong-to-right to total-change ratios as flagging rules.

Related to the regression comparisons, Bishop et al. flagged students with linear regression as well as Poisson regression residuals greater than 1.96. The authors note that if the model residuals were normally distributed, about 2.5% of students would have residuals above 1.96. Based on linear regression, *more* than 2.5% of students were flagged. However, about 2.5% were flagged by Poisson regression. Additionally, “very few, if any students, were jointly flagged by both procedures” (p. 7).

The authors then compared the number of students flagged using residual analysis to the number flagged using basic wrong-to-right counts. While the procedures flagged about the same percentage of students, the two methods did not flag the same students. Only “about 50 percent of the students flagged using the Poisson residuals, were *not* flagged using the [wrong-to-right] count (and vice versa)” (p. 9). Based on this analysis, the authors conclude that “a regression

modeling approach can identify unique outliers compared with a univariate selection model using [wrong-to-right] counts” (p. 7).

Summary

A review of the literature related to answer-changing and response-time analysis revealed several points of consideration for further exploration. First, answer-changing analysis has been conducted in a variety of settings and contexts with researchers reporting varying results and exploring many sources of variance. However, since most studies were small scale and based on paper-pencil assessments, it is worthwhile to explore those sources of variance in the large-scale, computer-based assessment context to expand the field’s knowledge in this area.

Additionally, several statistical models have been applied to answer-changing analysis. Exploration can expand on previous work by fitting models to computer-based assessment data. A study that builds on previous results found in the literature is described in the remaining chapters.

CHAPTER THREE - METHODS

This chapter is comprised of three sections: (1) a summary of the purpose and research questions, (2) a description of the sources of data for the study, and (3) a description of the data analysis methods employed.

Purpose Overview and Research Questions

While a number of studies have described answer changing in various settings, with various age-groups, and with tests of various types, all documented the construct in terms of paper-pencil testing, using either hand or optical scanner detection of changed item responses. However, there is much to learn about this evolving construct. The purpose of this study was to explore and document answer-changing patterns of students grades 3-11 on computer-based mathematics and English language arts mandated state achievement tests.

The following questions were addressed:

1. How can frequent answer-changing on computer-based achievement tests be understood?
2. How do student and item characteristics relate to answer-changing on computer-based achievement tests?
3. What statistical models and distributions are appropriate for answer-changing variables and analysis?

Data Source and Participants

Student performance on one Midwestern state's 2011-2012 summative achievement assessments in English language arts and mathematics were examined. The state administered almost 100 percent of the summative tests via computer-based assessment software provided by the state. The software allowed students to answer items in any order, as well as to review and

change answers as frequently as desired. The test was not timed. Student responses were collected, stored electronically, and maintained by the state's test vendor.

Assessment Forms and Administration

The assessments were administered in third through eleventh grades. However, in this particular testing year, mathematics was only tested at the 10th grade level and English language arts was only tested at the 11th grade level.

The number of items on each form ranged from 54 to 86 across grades and subjects, with some forms containing embedded field test items. For the purposes of this study, only items which were common across forms were examined, thus field test items were removed from study. All items were multiple choice with four possible answer choices per item.

The assessments were available to students in grades third through eighth grade from mid-February to mid-April 2012. Students in grades ten and eleven completed the assessments between October 2011 and mid-January 2012. For all grade levels, both the English language arts and mathematics assessments were administered in three parts. The tests were not timed, however, the suggested duration of test administration was 45-60 minutes per section.

Finally, it should be noted that in this state, the summative assessments were not used in teacher evaluation or as a requirement for graduation. Thus, there was limited motivation overall to cheat or influence student test scores.

Response-History Logs

In order to capture answer changing, the computer-based assessment system logged each student's path through the test, noting the answer marked on the screen when the student navigated away from the item, as well as the time spent viewing the screen. Thus, the system collected how many times the student reviewed an item, as well as any changes the student made

to the item and the response time. Additionally, the testing system allowed examination of precise answer-changing patterns, for example, whether a student changed an item one, two, or even ten times. The computer logs were the primary source of information for this study.

Participants

The students in this sample represented the entire population of students taking online summative achievement assessments in one Midwestern state. The number of students taking annual summative assessments in English language arts and mathematics in 2011-2012 were 252,736 and 255,984 respectively. The students were from 307 public districts and 1,390 public buildings. A number of private school students also took assessments. Table 1 shows the number of students who completed English language arts and mathematics assessments per grade. In this particular testing year, English language arts was not tested at the 10th grade level and mathematics was not tested at the 11th grade level. Population characteristics are displayed in Table 2.

Table 1

Number of Students Completing Summative Assessments in 2011-2012

Grade	English language arts	Mathematics
3	36,425	36,697
4	35,735	35,942
5	36,094	36,182
6	35,825	35,863
7	35,373	35,435
8	33,792	33,668
10	0	42,197
11	39,494	0
Total	252,736	255,984

Table 2

Characteristics of Examinees as a Percentage of Total Population

Group	State Percentage	
	ELA	Math
Race Ethnicity		
Asian	2.63	2.66
African American	6.50	6.49
Hispanic	16.47	16.50
Multicultural or Missing	4.31	4.29
Native American	1.11	1.11
Pacific Islander	.15	.15
White	68.76	68.75
Economically disadvantaged	44.82	44.79
Other		
English Language Learners	9.65	9.67
Female	49.09	49.06
Male	50.85	50.92
Students with disabilities	9.23	9.45

Economically disadvantaged students were defined by the state as students who received free or reduced lunches through the National School Lunch Program.

English language learners were students who were served through the Title III program of the Elementary and Secondary Education Act (ESEA) or received services through limited English proficiency programs.

Students with Disabilities were defined as:

. . . children with intellectual disability, hearing impairment including deafness, speech or language impairment, visual impairment including blindness, serious emotional disturbance, orthopedic impairment, autism, traumatic brain injury, developmental delay, other health impairment, specific learning disability, deaf-blindness, or multiple

disabilities, and who, by reason thereof, receive special education and related services under the Individuals with Disabilities Education Act (IDEA) according to an individualized education program (IEP), individual family service plan (IFSP), or a services plan provided under IDEA. (United States Department of Education, 2012, p. 3)

Procedure

From the answer-changing logs and item-response data, a variety of exploratory analyses were conducted. For all approaches, a wrong answer change or a right answer change was defined via the final change made to an item, regardless of how the student had answered the item initially.

Descriptive Analysis Procedure

Descriptive information about answer-changing patterns including total changes, wrong-to-right changes, and right-to-wrong changes were recorded at both the student level and item level. Descriptive information at the student level included frequency distributions, state-level means, and standard deviations. Additionally, contingency tables comparing wrong-to-right to total changes, right-to-wrong to total changes, and right-to-wrong to wrong-to-right changes were constructed. Also at the student level, basic answer-changing patterns for three, four, or five item changes per student were documented.

Related to response time, the distributions of average screen time per wrong-to-right and right-to-wrong changes were calculated for each content area, including means and standard deviations. Additionally, the distributions of wrong-to-right changes in less than 60 seconds were listed, as well as the number of wrong-to-right changes per student in less than 10 seconds.

Modeling Procedure

Based on the distribution of total answer changes and wrong-to-right answer changes at the student level, the fit of Poisson and negative binomial regression models were explored.

Based on the thorough review of the literature in Chapter 2, an a priori global model was specified which included potentially relevant effects (Burnham & Anderson, 2002) for each dependent variable, total changes, wrong-to-right changes, and wrong-to-right to total changes proportion. No interaction effects were included in the global model. The predictors in the global model included:

- grade level
- gender
- race
- whether or not the student received free or reduced lunches (lunch)
- whether or not the student received English as second language (ESOL) services
- whether or not the student received special education services (sped)

Based on the global model, a pool of a priori, alternate models was also created using combinations of the same predictors listed above, with some models including interaction effects and others excluding certain predictors entirely. Creating a pool of alternate models allowed for exploration of which combinations of variables would provide evidence of parsimony. The pool of candidate model is listed in Table 3 below, with “+” indicating a model with only main effects and “*” indicating a model with interaction effects among the parameters.

Table 3

Candidate Models

Model Number	Predictors
1 (Global)	grade + gender + race + lunch + ESOL + sped
2	grade + gender + race + lunch + ESOL
3	grade + gender + race + lunch
4	grade + race + lunch + ESOL
5	grade * race * lunch * ESOL
6	grade * gender * race * lunch * ESOL
7	gender * race * lunch * ESOL
8	gender + race + lunch + ESOL
9	lunch + ESOL
10	ESOL
11	race + lunch + ESOL

For the dependent variables total changes and wrong-to-right changes, each model was applied to the English language arts and mathematics data separately using Poisson and negative binomial regression. In an additional step, each model was applied to restricted English language arts and mathematics data sets that only included students who had made one or more answer changes.

For the dependent variable wrong-to-right changes to total changes proportion, the proportion represented a rate of answer changing behavior. To model this rate, each model was fit using an offset variable which was the log of the count of total item changes (Coxe et al., 2009). These models were applied only to the restricted data set of students who had one or more answer changes.

Model Selection

An information theoretic approach was used to select the most parsimonious model based on each model's Akaike information criterion (AIC). In general, the AIC represents the amount of information lost when a particular model is chosen to represent the true phenomenon that

generated the data (Mazerolle, 2004). The AIC is calculated from the log-likelihood of the model given the data and the number of parameters in the model (K) (Burnham & Anderson, 2002).

The AIC is defined as:

$$AIC = -2(\log\text{-likelihood}) + 2K \quad (3)$$

The model with the smallest AIC can be considered as a “best” model. Additionally, the AIC increases as the number of estimated parameters (K) increases. AIC only allows models to be compared to each other; it does not prevent one from choosing a poor quality, or poorly fitting model.

Model Comparisons

Candidate models were compared via two values, delta AIC and Akaike weights. Delta AIC represents the difference in AICs between the target model and the best model. Delta AIC is defined as:

$$\text{Delta AIC} = \Delta_i = AIC_i - \min AIC \quad (4)$$

Rules of thumb exist for using Δ_i to determine relative support for each model. In general, Δ_i of less than two suggests provides strong evidence for the model, whereas a Δ_i of greater than ten suggests that a model is not likely (Burnham & Anderson, 2002).

Akaike weights provide another measure of the strength of evidence for each model relative to the pool of possible models. Specifically, Akaike weights represent the probability that a particular model is the best model in the pool. For example, an Akaike weight of .50 can be interpreted as one model having a 50% chance of being the best model in the pool.

$$\text{Akaike weight} = w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta_r)} \quad (5)$$

The weights can also be used to compare models by calculating an evidence ratio of the Akaike weight of the best model to the Akaike weight of a competing models. The resulting value allows one to determine how many more times likely the best model is compared to the competing model (Mazerolle, 2004).

Item-level Analysis Procedure

An exploration of the data with the item as the unit of analysis was completed. Descriptive statistics including the wrong-to-right changes per item and the right-to-wrong changes per item were calculated and reported. Additionally, the relationships between item difficulty (p-values) and total answer changes as well as the relationships between item difficulty and wrong-to-right answer changes were plotted.

Summary

This chapter presented the methods used to address the study's research questions including data sources, demographics of participants, and data analysis procedures.

CHAPTER FOUR - RESULTS

The purpose of this study was to explore and document answer-changing patterns of students grades 3-12 on computer-based English language arts and mathematics mandated state achievement tests. The first part of this chapter presents the distributions of answer-changing patterns at the student level. Next, the chapter presents the results of statistical modeling of answer-changing variables. The chapter concludes with the results of an item-level analysis of the answer-changing data.

Answer-Changing Descriptive Statistics and Distributions

Thirty percent of students taking English language arts assessments changed at least one item; 1% changed five or more items. Fifty-seven percent of students taking mathematics assessments changed at least one item; 5% changed five or more items. In English language arts, students generally changed fewer items overall, including wrong-to-right item changes. The overall state mean of answer changes per student in English language arts was .52, (*SD* 1.31). Table 4 summarizes general answer-changing behavior.

Table 4

Statewide Student Answer-Changing Descriptive Statistics

	English language arts	Mathematics
Percent changing at least one item	30%	57%
Percent changing five or more items	1%	5%
Mean, total answers changed	.52	1.31
Standard deviation, total answers changed	1.13	2.04
Mean, wrong-to-right changes	.27	.78
Standard deviation wrong-to-right changes	.67	1.35
Mean, right-to-wrong changes	.12	.21
Standard deviation, right-to-wrong changes	.39	.53

Total Answer-Change Frequencies

Distributions of total answer changes are shown in Table 5 for English language arts and Table 6 for mathematics. Distributions for total-answer changes by grade-level were found to be extremely similar to overall distributions for both English language arts and mathematics and thus are not displayed. Total answer-changing is positively skewed for both content areas. English language arts had skewness of 6.97 (*SE* 0.005) and kurtosis of 172.56 (*SE* 0.99). Math had skewness of 4.63 (*SE* 0.005) and kurtosis of 58.14 (*SE* 0.99).

Table 5

Frequencies of Total Answer Changes – English Language Arts

Total Answer Changes	Count of Students	Percent	Cumulative Percent
0	176,739	69.93	69.93
1	48,277	19.10	89.03
2	15,864	6.28	95.31
3	6,125	2.42	97.73
4	2,655	1.05	98.78
5	1,288	.51	99.29
6	691	.27	99.57
7	372	.15	99.71
8	223	.09	99.80
9	149	.06	99.86
10	81	.03	99.89
11	75	.03	99.92
12	50	.02	99.94
13	40	.02	99.96
14	24	.01	99.97
15	16	.01	99.97
16	13	.01	99.98
17	7	.00	99.98
18	11	.00	99.99
19	2	.00	99.99
20	5	.00	99.99
21	1	.00	99.99
22	4	.00	99.99
23	5	.00	99.99
24	3	.00	99.99
25	2	.00	99.99
26	2	.00	100.00
27	2	.00	100.00
29	2	.00	100.00
30	2	.00	100.00
36	1	.00	100.00
37	1	.00	100.00
41	1	.00	100.00

Total Answer Changes	Count of Students	Percent	Cumulative Percent
48	1	.00	100.00
67	1	.00	100.00
68	1	.00	100.00

Table 6

Frequencies of Total Answer Changes - Mathematics

Total Answer Changes	Count of Students	Percent	Cumulative Percent
0	110,855	43.31	43.31
1	67,897	26.52	69.83
2	35,150	13.73	83.56
3	18,076	7.06	90.62
4	9,563	3.74	94.36
5	5,361	2.09	96.45
6	3,020	1.18	97.63
7	1,860	.73	98.36
8	1,181	.46	98.82
9	718	.28	99.10
10	556	.22	99.32
11	414	.16	99.48
12	280	.11	99.59
13	187	.07	99.66
14	150	.06	99.72
15	150	.06	99.78
16	102	.04	99.82
17	80	.03	99.85
18	67	.03	99.88
19	59	.02	99.90
20	48	.02	99.92
21	37	.01	99.93
22	45	.02	99.95
23	25	.01	99.96
24	14	.01	99.97
25	15	.01	99.97
26	10	.00	99.97
27	14	.01	99.98
28	2	.00	99.98
29	6	.00	99.98
30	4	.00	99.99
31	11	.00	99.99
32	5	.00	99.99
33	1	.00	99.99
34	2	.00	99.99
35	2	.00	99.99
36	1	.00	99.99
37	1	.00	99.99
38	2	.00	99.99
40	3	.00	100.00
41	2	.00	100.00
42	1	.00	100.00
44	1	.00	100.00

Total Answer Changes	Count of Students	Percent	Cumulative Percent
47	1	.00	100.00
48	1	.00	100.00
53	1	.00	100.00
59	1	.00	100.00
76	1	.00	100.00
85	1	.00	100.00

Wrong-to-Right Frequencies

Wrong-to-right answer change frequencies are shown in Table 7 for English language arts and Table 8 for mathematics. The overall state average for wrong-to-right changes in English language arts was .27 (*SD* 0.67). The state average for mathematics wrong-to-right changes was .78, (*SD* 1.35). Wrong-to-right answer changes are positively skewed for both content areas. English language arts had skewness of 4.75 (*SE* 0.005) and kurtosis of 56.48 (*SE* 0.99). Math had skewness of 4.67 (*SE* 0.005) and kurtosis of 48.00 (*SE* 0.99).

Table 7

Wrong-to-Right Frequencies – English Language Arts

Wrong-To-Right Changes	Count of Students	Percent	Cumulative Percent
0	202,926	80.29	80.29
1	37,707	14.92	95.21
2	8,387	3.32	98.53
3	2,314	.92	99.45
4	765	.30	99.75
5	320	.13	99.87
6	145	.06	99.93
7	78	.03	99.96
8	26	.01	99.97
9	30	.01	99.98
10	9	.00	99.99
11	11	.00	99.99
12	2	.00	99.99
13	1	.00	99.99
14	5	.00	100.00
16	3	.00	100.00
17	3	.00	100.00
18	1	.00	100.00
21	1	.00	100.00
24	2	.00	100.00

Table 8

Wrong-to-Right Frequencies – Mathematics

Wrong-To-Right Changes	Count of Students	Percent	Cumulative Percent
0	143,540	56.07	56.07
1	67,673	26.44	82.51
2	25,965	10.14	92.65
3	10,034	3.92	96.57
4	4,078	1.59	98.17
5	1,873	.73	98.90
6	925	.36	99.26
7	531	.21	99.47
8	348	.14	99.60
9	254	.10	99.70
10	196	.08	99.78
11	114	.04	99.82
12	100	.04	99.86
13	79	.03	99.89
14	60	.02	99.92
15	57	.02	99.94
16	24	.01	99.95
17	31	.01	99.96
18	24	.01	99.97
19	11	.00	99.97
20	17	.01	99.98
21	12	.00	99.99
22	9	.00	99.99
23	5	.00	99.99
24	3	.00	99.99
25	8	.00	99.99
26	6	.00	100.00
27	1	.00	100.00
28	1	.00	100.00
29	3	.00	100.00
33	1	.00	100.00
38	1	.00	100.00

Right-to-Wrong Frequencies

Right-to-wrong frequencies are shown in Table 9 for English language arts and Table 10 for mathematics. The average count of right-to-wrong item changes across the state in English language arts was .12 (*SD* 0.39). The average count of right-to-wrong for mathematics was .21 (*SD* 0.53). Overall, there were fewer right-to-wrong changes than wrong-to-right changes.

Additionally, English language arts had skewness of 5.44 (*SE* 0.005) and kurtosis of 81.39 (*SE* 0.99). Math had skewness of 4.52 (*SE* 0.005) and kurtosis of 51.23 (*SE* 0.99).

Table 9

Right-to-Wrong Frequencies – English Language Arts

Right-To-Wrong Changes	Count of Students	Percent	Cumulative Percent
0	226,287	89.53	89.53
1	23,012	9.11	98.64
2	2,821	1.11	99.76
3	477	.19	99.94
4	89	.03	99.98
5	29	.01	99.99
6	6	.00	99.99
7	11	.00	99.99
8	1	.00	99.99
9	2	.00	99.99
10	0	.00	99.99
11	2	.00	99.99
12	1	.00	100.00

Table 10

Right-to-Wrong Frequencies – Mathematics

Right-To-Wrong Changes	Count of Students	Percent	Cumulative Percent
0	213,063	83.23	83.23
1	34,763	13.58	96.81
2	6,344	2.48	99.29
3	1,287	.50	99.79
4	336	.13	99.93
5	116	.33	99.97
6	46	.04	99.99
7	12	.01	99.99
8	11	.00	99.99
9	2	.00	99.99
10	3	.00	99.99
11	0	.00	99.99
12	0	.00	99.99
13	1	.00	100.00

Contingency Tables

Contingency tables show an interesting view of the frequency information. Table 11 shows wrong-to-right counts versus total items changes for English language arts. Table 12

shows the same information for mathematics. The diagonals show a 1:1 proportion of wrong-to-right changes to total changes. The values on the diagonals are slightly higher for mathematics, reflecting more overall changes and wrong-to-right changes per student than in English language arts. The values on the diagonal fall off considerably after 16 changes for English language arts and after 11 changes for mathematics. The lower right quadrant of the table also shows students who had a high proportion of wrong-to-right to total changes. These students might be worth further examination.

Table 11

Changes from Wrong-to-Right by Total Item Changes – English Language Arts

Total Answer Changes	Wrong-to-Right Answer Changes																			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	21	24
0	176,739	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	22,424	28,235	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3,045	7,133	5,212	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	546	1,685	2,070	1,110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	128	442	695	700	289	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	31	138	244	272	225	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	6	45	100	121	114	103	39	0	0	0	0	0	0	0	0	0	0	0	0	0
7	4	13	34	49	64	52	29	16	0	0	0	0	0	0	0	0	0	0	0	0
8	1	9	11	21	38	36	25	14	0	0	0	0	0	0	0	0	0	0	0	0
9	1	3	7	19	14	15	17	18	5	6	0	0	0	0	0	0	0	0	0	0
10	0	1	7	9	9	9	12	8	2	7	0	0	0	0	0	0	0	0	0	0
11	1	1	4	5	6	6	10	5	8	7	3	1	0	0	0	0	0	0	0	0
12	0	1	0	1	4	2	3	7	4	3	3	4	0	0	0	0	0	0	0	0
13	0	0	2	1	0	1	2	4	0	1	3	3	0	0	0	0	0	0	0	0
14	0	0	1	0	0	0	3	0	0	2	0	1	1	0	1	0	0	0	0	0
15	0	0	0	1	2	0	2	0	2	0	0	1	0	0	0	0	0	0	0	0
16	0	1	0	2	0	0	1	1	1	3	0	0	0	0	2	0	0	0	0	0
17	0	0	0	0	0	2	1	3	1	1	0	0	1	0	0	0	0	0	0	0
18	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1	1	1	0	0	0
19	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
23	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0
25	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2

Total Answer Changes	Wrong-to-Right Answer Changes																			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	21	24
30	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
34	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Note. Total Students = 252,736

Table 12

Changes from Wrong-to-Right by Total Item Changes – Mathematics

Total Answer Changes	Wrong-to-Right Answer Changes																																
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	33	38	
0	110,855	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	24,626	45,866	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	5,862	14,849	15,136	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1,545	4,658	6,544	5,066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	427	1,463	2,602	2,631	1,700	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	139	497	977	1,267	1,133	693	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	52	185	377	546	606	533	311	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	17	80	163	249	287	277	249	165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	8	33	82	131	151	151	145	144	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3	18	27	54	81	91	81	79	95	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	16	25	34	50	45	49	50	58	64	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	2	3	7	21	17	25	30	39	35	40	52	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	3	8	15	10	13	18	18	19	38	37	25	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	2	2	5	6	9	15	11	9	8	21	21	20	27	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	2	6	11	9	8	7	12	9	11	6	21	15	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	4	3	6	8	4	7	6	5	10	6	10	15	16	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	3	2	5	4	2	1	3	6	7	7	8	10	15	17	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	1	1	4	1	3	4	4	6	4	0	7	6	3	6	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	1	2	0	1	2	3	2	1	2	4	4	4	5	6	9	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	1	4	1	2	1	0	3	4	2	0	2	5	3	7	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	1	4	1	4	1	0	1	1	0	1	1	2	3	0	8	5	4	6	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	1	0	0	2	3	0	0	1	0	0	0	3	1	0	0	2	4	2	3	2	0	0	0	0	0	0	0	0	0	0	0
22	0	0	1	0	0	0	2	1	1	0	0	1	2	2	2	2	1	1	4	1	3	4	1	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	0	1	1	0	0	4	1	0	0	0	0	0	0	0	0	0

Total		Wrong-to-Right Answer Changes																																		
Answer Changes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	33	38				
24	0	0	0	0	1	0	1	0	0	0	0	1	0	2	0	1	2	0	1	0	1	2	1	1	0	0	0	0	0	0	0	0	0			
25	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	3	3	1	2	0	0	0	0	0	0	0			
26	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	2	0	0	2	0	0	0	0	0	0			
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	1	0	0	0	0	0			
28	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
29	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0		
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	2	0	0	0	
32	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Note. Total Students = 255,984

In contrast, Tables 13 and 14 show right-to-wrong changes versus total item changes for English language arts and mathematics respectively. The one-to-one proportion of right-to-wrong changes to total changes falls off quickly after two changes.

Table 13

Right-to-Wrong Frequencies Versus Total Item Changes – English Language Arts

Total Answer Changes	Right-to-Wrong Answer Changes													
	0	1	2	3	4	5	6	7	8	9	10	11	12	
0	176,739	0	0	0	0	0	0	0	0	0	0	0	0	
1	37,052	13,607	0	0	0	0	0	0	0	0	0	0	0	
2	8,648	5,679	1,063	0	0	0	0	0	0	0	0	0	0	
3	2,396	2,131	775	110	0	0	0	0	0	0	0	0	0	
4	831	857	455	101	10	0	0	0	0	0	0	0	0	
5	314	368	230	77	11	3	0	0	0	0	0	0	0	
6	149	171	124	67	16	1	0	0	0	0	0	0	0	
7	68	78	60	44	8	3	0	0	0	0	0	0	0	
8	24	50	41	29	8	2	0	1	0	0	0	0	0	
9	24	31	23	13	7	4	1	1	0	1	0	0	0	
10	16	10	13	13	8	3	0	1	0	0	0	0	0	
11	6	11	21	8	4	4	0	3	0	0	0	0	0	
12	7	11	4	6	3	0	0	1	0	0	0	0	0	
13	3	3	4	3	2	2	0	0	0	0	0	0	0	
14	3	2	1	1	1	1	0	0	0	0	0	0	0	
15	0	0	3	2	2	0	1	0	0	0	0	0	0	
16	1	0	0	0	3	3	2	0	0	0	0	0	0	
17	2	0	0	0	3	0	2	1	1	0	0	0	0	
18	2	0	1	0	2	0	0	0	0	0	0	1	0	
19	0	1	0	0	0	0	0	2	1	0	0	0	0	
20	0	0	0	0	1	0	0	0	0	1	0	0	0	
22	0	0	0	1	0	0	0	0	0	0	0	0	0	
23	0	0	2	0	0	0	0	0	0	0	0	0	0	
25	0	0	0	0	0	2	0	1	0	0	0	0	0	
26	1	0	1	2	0	0	0	0	0	0	0	0	0	
30	0	0	0	0	0	0	0	0	0	0	0	1	0	
31	0	0	0	0	0	1	0	0	0	0	0	0	0	
34	0	0	0	0	0	0	0	0	0	0	0	0	1	

Note. Total Students=252,736

Table 14

Right-to-Wrong Frequencies Versus Total Item Changes – Mathematics

Total Answer Changes	Right-to-Wrong Answer Changes												
	0	1	2	3	4	5	6	7	8	9	10	13	
0	110,855	0	0	0	0	0	0	0	0	0	0	0	
1	58,683	11,809	0	0	0	0	0	0	0	0	0	0	
2	24,553	9,986	1,308	0	0	0	0	0	0	0	0	0	
3	10,199	5,893	1,554	167	0	0	0	0	0	0	0	0	
4	4,164	3,149	1,259	231	20	0	0	0	0	0	0	0	
5	1,979	1,664	809	216	36	2	0	0	0	0	0	0	
6	981	898	525	168	30	7	1	0	0	0	0	0	
7	531	468	291	143	45	9	0	0	0	0	0	0	
8	312	276	212	90	43	12	1	0	0	0	0	0	
9	208	173	106	62	26	12	1	0	0	0	0	0	
10	146	110	79	58	32	8	4	1	1	0	0	0	
11	112	76	48	38	21	11	3	1	0	0	0	0	
12	65	64	48	20	9	9	6	0	1	0	0	0	
13	58	43	28	23	12	7	4	0	0	0	1	0	
14	40	33	19	17	13	4	4	0	1	0	0	0	
15	38	29	17	7	11	6	5	1	0	0	0	0	
16	30	23	9	9	12	7	2	1	1	0	0	0	
17	17	12	6	6	8	7	3	0	0	0	0	0	
18	16	14	7	9	4	0	2	0	0	0	0	0	
19	12	11	7	2	1	4	2	1	0	0	0	0	
20	19	6	3	6	4	2	2	0	0	1	0	0	
21	11	3	2	2	2	0	3	0	1	0	0	0	
22	8	7	3	4	2	3	2	0	0	0	0	0	
23	6	2	0	1	2	0	0	0	0	0	0	0	
24	3	3	0	1	3	1	1	2	0	0	0	0	
25	3	4	3	0	0	2	0	1	1	0	0	0	
26	4	1	0	0	0	2	0	0	1	0	0	0	
27	6	3	0	0	0	0	0	0	0	0	0	0	
28	0	0	0	0	0	0	0	2	1	0	0	0	
29	0	0	1	0	0	0	0	1	1	0	0	0	
30	0	2	0	1	0	0	0	0	1	0	0	0	
31	3	1	0	1	0	0	0	0	0	0	0	0	
32	1	0	0	0	0	0	0	0	1	0	0	0	
33	0	0	0	1	0	0	0	0	0	0	0	0	
34	6	0	0	2	0	1	0	0	0	0	0	0	
36	0	0	0	1	0	0	0	0	0	0	0	0	
37	0	0	0	0	0	0	0	0	0	0	1	0	
39	0	0	0	0	0	0	0	0	0	0	0	1	
40	0	0	0	0	0	0	0	1	0	0	0	0	

42	0	0	0	0	0	0	0	0	0	0	1	0
45	0	0	0	1	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	1	0	0

Note. Total Students = 252,736

Table 15 presents another interesting view of this data, a depiction of right-to-wrong changes to wrong-to-right changes for English language arts. Table 16 presents the same information for mathematics. In the context of test-score integrity, one might be less concerned with students who have a high number of both types. Students who have high numbers of wrong-to-right but low numbers of right-to-wrong changes may be worth closer evaluation.

Table 15

Right-to-Wrong Frequencies Versus Wrong-to-Right Frequencies – English Language Arts

Wrong-to-Right Answer Changes	Right-to-Wrong Answer Changes												
	0	1	2	3	4	5	6	7	8	9	10	11	12
0	186,216	15,207	1,337	143	17	3	0	1	0	1	0	0	0
1	31,514	5,217	818	138	14	5	1	1	0	0	0	0	0
2	6,272	1,640	367	81	18	7	0	2	0	0	0	0	0
3	1,529	570	140	52	11	5	2	4	0	1	0	0	0
4	447	193	81	34	9	0	0	1	0	0	0	0	0
5	167	101	30	12	6	2	0	0	1	0	0	1	0
6	74	40	17	6	5	1	1	0	0	0	0	1	0
7	28	24	11	7	4	1	1	1	0	0	0	0	1
8	8	3	8	1	3	1	1	1	0	0	0	0	0
9	14	7	6	0	1	2	0	0	0	0	0	0	0
10	5	4	0	0	0	1	0	0	0	0	0	0	0
11	4	5	2	0	0	0	0	0	0	0	0	0	0
12	2	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	1	0	0	0	0	0	0	0	0
14	2	1	2	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	1	0	1	0	1	0	0	0	0	0	0	0
17	0	9	1	1	0	0	0	0	0	0	0	0	0
18	0	2	0	1	0	0	0	0	0	0	0	0	0
19	0	4	0	0	0	0	0	0	0	0	0	0	0
20	0	3	0	0	0	0	0	0	0	0	0	0	0
21	0	4	0	0	0	1	0	0	0	0	0	0	0
22	0	4	0	0	0	0	0	0	0	0	0	0	0
23	0	1	0	0	0	0	0	0	0	0	0	0	0

Wrong-to-Right Answer Changes	Right-to-Wrong Answer Changes												
	0	1	2	3	4	5	6	7	8	9	10	11	12
24	1	2	1	1	0	0	0	0	0	0	0	0	0

Table 16

Right-to-Wrong Frequencies Versus Wrong-to-Right Frequencies – Mathematics

Wrong-to-Right	Right-to-Wrong Answer Changes												
Answer Changes	0	1	2	3	4	5	6	7	8	9	10	13	
0	125,776	15,407	2,000	301	48	4	3	1	0	0	0	0	
1	55,221	10,222	1,830	312	62	19	6	1	0	0	0	0	
2	19,676	4,779	1,171	254	58	17	5	1	3	0	1	0	
3	7,012	2,161	609	168	57	22	5	0	0	0	0	0	
4	2,612	996	332	80	33	13	8	2	2	0	0	0	
5	1,144	498	137	50	26	12	4	0	1	1	0	0	
6	539	229	90	28	18	11	8	1	1	0	0	0	
7	308	126	48	31	8	8	1	0	1	0	0	0	
8	218	73	28	12	11	1	3	0	1	0	1	0	
9	135	66	33	10	4	4	1	1	0	0	0	0	
10	102	56	22	9	4	1	0	1	1	0	0	0	
11	70	24	11	5	0	1	0	1	0	0	1	1	
12	52	31	8	6	1	1	1	0	0	0	0	0	
13	39	23	9	3	2	1	1	1	0	0	0	0	
14	32	20	5	3	0	0	0	0	0	0	0	0	
15	33	12	4	4	1	1	0	2	0	0	0	0	
16	10	9	0	2	2	0	0	0	1	0	0	0	
17	21	8	2	0	0	0	0	0	0	0	0	0	
18	18	3	0	2	1	0	0	0	0	0	0	0	
19	7	3	1	0	0	0	0	0	0	0	0	0	
20	10	5	0	2	0	0	0	0	0	0	0	0	
21	7	2	2	1	0	0	0	0	0	0	0	0	
22	6	3	0	0	0	0	0	0	0	0	0	0	
23	3	1	1	0	0	0	0	0	0	0	0	0	
24	2	1	0	0	0	0	0	0	0	0	0	0	
25	4	3	1	0	0	0	0	0	0	0	0	0	
26	3	1	0	2	0	0	0	0	0	0	0	0	
27	1	0	0	0	0	0	0	0	0	0	0	0	
28	0	0	0	1	0	0	0	0	0	0	0	0	
29	2	1	0	0	0	0	0	0	0	0	0	0	
33	0	0	0	0	0	0	0	0	0	1	0	0	
38	0	0	0	1	0	0	0	0	0	0	0	0	

Answer-Changing Patterns

Table 17 shows the number of students who demonstrated various patterns of answer changes, up to four total changes, as revealed by the computer-log data. Very few students change their answers more than four times.

Table 17

Patterns of Answer Changes

Pattern	English language arts		Mathematics	
	Number of Students	Number of Items	Number of Students	Number of Items
Items Changed 1x				
WR*	47,743	64,132	109,723	190,231
RW	25,429	29,237	41,036	50,549
WW	17,941	21,037	45,418	60,793
Items Changed 2x				
RWR*	2,759	2,939	5,250	5,609
WWR*	1,064	1,107	3,093	3,305
RWW	492	496	834	865
W ₁ RW ₁ (same wrong)	877	893	1,872	1,925
W ₁ RW ₂ (different wrong)	439	449	911	941
W ₁ W ₂ W ₁ (same wrong)	655	676	1,824	1,888
W ₁ W ₂ W ₃ (different wrong)	247	256	594	608
Items Changed 3x				
WRWR*	132	132	424	435
WWWR*	62	62	193	196
RWRW	102	110	201	202
WRWW	39	39	76	76
WWWW	55	55	157	159
Items Changed 4x				
RWWWR*	13	13	18	18
WRWWR*	5	5	4	4
WWWRW	4	5	11	11

Note. * denotes a wrong-to-right pattern

Response Time

Response time represented the amount of time a student spent on a screen before surfing away to a different item. The average amount of total screen time per item was calculated for

English language arts and mathematics items across all students. For English language arts, the average response time per item ranged from 4.95 seconds to 154.75 seconds, with a mean of 28.45 seconds per item (*SD* 15.61). For mathematics, the average response times ranged from 8.70 seconds to 184.42 seconds, with a mean of 47.12 seconds (*SD* 24.79).

Average screen times for items changed from wrong to right were also calculated. For English language arts, the average wrong-to-right screen times ranged from 1.50 seconds to 209.86 seconds, with a mean of 28.50 (*SD* 17.32). For mathematics, the average wrong-to-right screen times ranged from 8.02 seconds to 214.36 seconds, with a mean of 49.23 seconds (*SD* 27.59). The distributions of average screen time for wrong-to-right changes for English language arts and mathematics are shown in Figure 1.

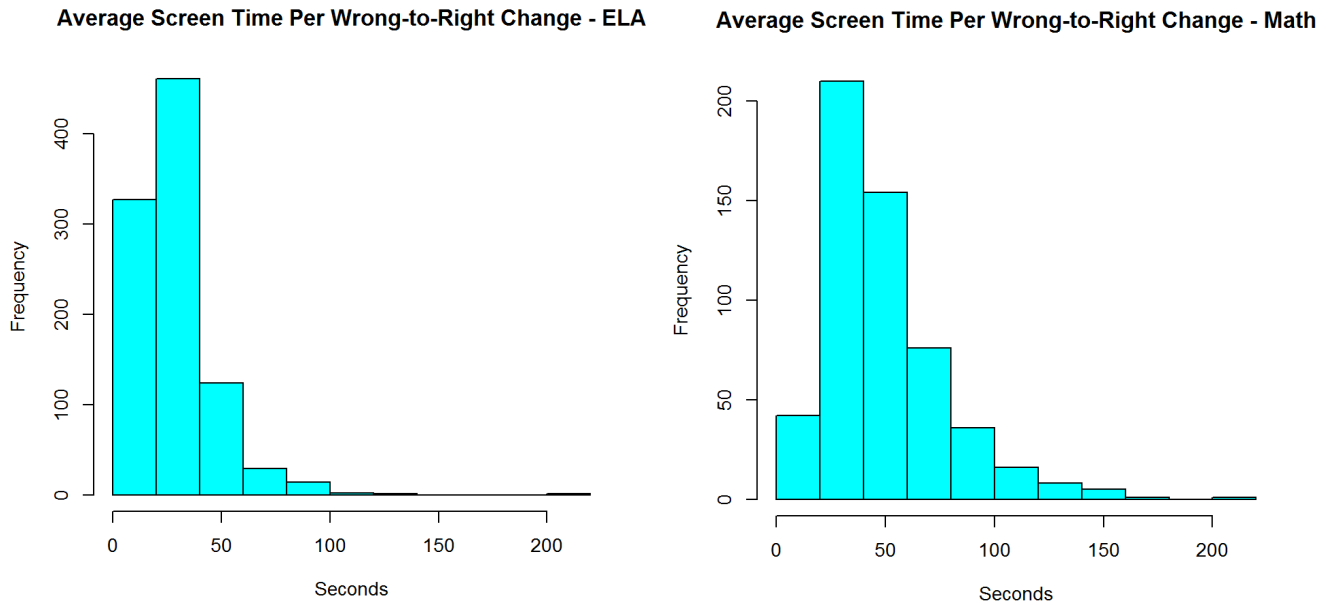


Figure 1. Distribution of average screen time per wrong-to-right change – English language arts and mathematics

The average screen times for items changed from right-to-wrong were also calculated. For English language arts, the average screen times ranged from 1 second to 635 seconds, with a mean of 28.39 (*SD* 35.82). For mathematics, the average screen times ranged from 9.92 to 154.45 seconds, with a mean of 43.15 seconds (*SD* 23.07). The distributions of screen times for items changed from right to wrong are shown in Figure 2.

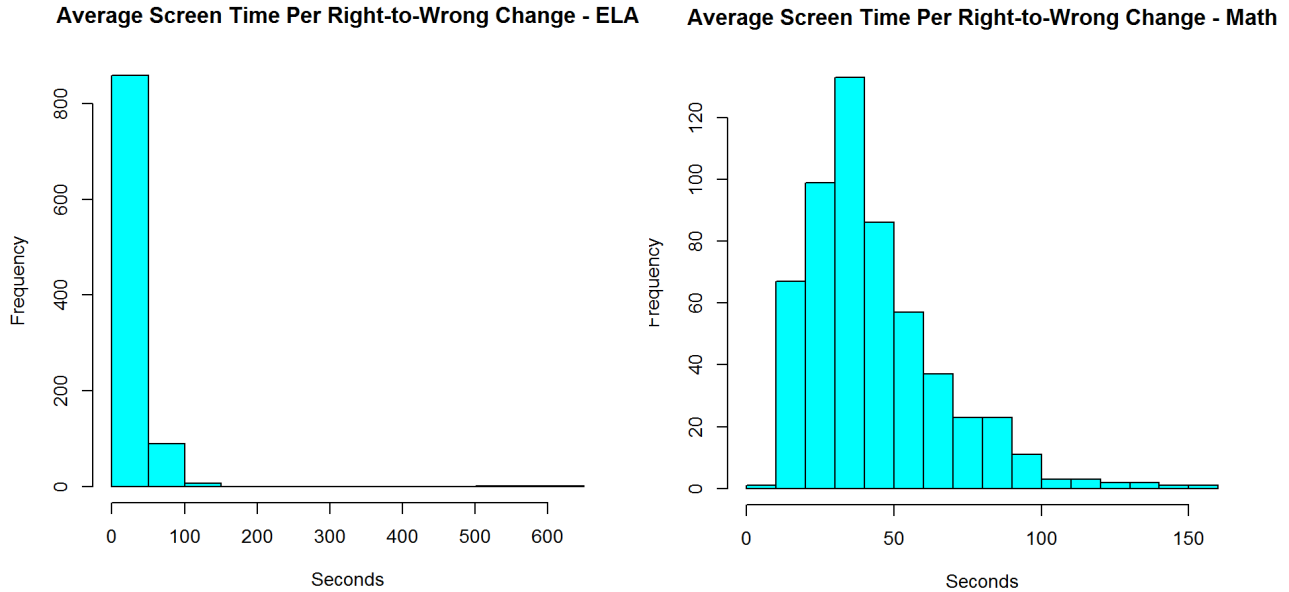


Figure 2. Distribution of average screen time per right-to-wrong change – English language arts and mathematics

The distributions of response times for wrong-to-right answer changes made in less than 60 seconds are listed in Table 18 for English language arts and Table 19 for mathematics. Note that there are larger numbers of wrong-to-right changes at the smaller time points than larger time points.

Table 18

Frequencies of Wrong-to-Right Item Changes 60 Seconds or Less Elapsed Screen Time – English Language Arts

Seconds	Wrong-to-Right Changes	Percent	Cumulative Percent
0	3,104	12.70	12.70
1	3,051	12.48	25.17
2	2,666	10.90	36.08
3	1,704	6.97	43.05
4	1,351	5.53	48.57
5	1,065	4.36	52.93
6	859	3.51	56.44
7	709	2.90	59.34
8	606	2.48	61.82
9	546	2.23	64.06
10	484	1.98	66.04
11	409	1.67	67.71
12	392	1.60	69.31

Seconds	Wrong-to-Right Changes	Percent	Cumulative Percent
13	382	1.56	70.87
14	337	1.38	72.25
15	332	1.36	73.61
16	298	1.22	74.83
17	300	1.23	76.06
18	264	1.08	77.14
19	260	1.06	78.20
20	241	.99	79.19
21	246	1.01	80.19
22	249	1.02	81.21
23	199	.81	82.02
24	232	.95	82.97
25	191	.78	83.75
26	186	.76	84.51
27	181	.74	85.26
28	197	.81	86.06
29	170	.70	86.76
30	166	.68	87.44
31	169	.69	88.13
32	163	.67	88.79
33	149	.61	89.40
34	142	.58	89.98
35	146	.60	90.58
36	141	.58	91.16
37	129	.53	91.68
38	127	.52	92.20
39	114	.47	92.67
40	102	.42	93.09
41	128	.52	93.61
42	90	.37	93.98
43	85	.35	94.33
44	105	.43	94.76
45	98	.40	95.16
46	101	.41	95.57
47	98	.40	95.97
48	80	.33	96.30
49	83	.34	96.64
50	98	.40	97.04
51	102	.42	97.46
52	69	.28	97.74
53	78	.32	98.06
54	59	.24	98.30
55	74	.30	98.60
56	86	.35	98.95
57	65	.27	99.22
58	72	.29	99.51
59	63	.26	99.77
60	56	.23	100.00

Table 19

Frequencies of Wrong-to-Right Item Changes 60 Seconds or Less Elapsed Screen Time – Mathematics

Seconds	Wrong-to-Right Changes	Percent	Cumulative Percent
1	110	.13	.13
2	3,043	3.61	3.74
3	5,245	6.22	9.96
4	5,026	5.96	15.91
5	4,344	5.15	21.06
6	3,603	4.27	25.33
7	3,202	3.80	29.13
8	2,794	3.31	32.44
9	2,543	3.01	35.46
10	2,291	2.72	38.17
11	2,198	2.61	40.78
12	2,107	2.50	43.27
13	1,980	2.35	45.62
14	1,901	2.25	47.88
15	1,847	2.19	50.06
16	1,660	1.97	52.03
17	1,633	1.94	53.97
18	1,590	1.88	55.85
19	1,526	1.81	57.66
20	1,556	1.84	59.51
21	1,482	1.76	61.26
22	1,406	1.67	62.93
23	1,325	1.57	64.50
24	1,320	1.56	66.07
25	1,270	1.51	67.57
26	1,218	1.44	69.01
27	1,180	1.40	70.41
28	1,141	1.35	71.77
29	1,104	1.31	73.07
30	1,065	1.26	74.34
31	1,077	1.28	75.61
32	980	1.16	76.78
33	951	1.13	77.90
34	941	1.12	79.02
35	926	1.10	80.12
36	921	1.09	81.21
37	900	1.07	82.27
38	893	1.06	83.33
39	827	0.98	84.31
40	755	0.89	85.21
41	824	0.98	86.19
42	739	0.88	87.06
43	708	0.84	87.90
44	730	0.87	88.77
45	706	0.84	89.60
46	651	0.77	90.37
47	682	0.81	91.18
48	643	0.76	91.95

Seconds	Wrong-to-Right Changes	Percent	Cumulative Percent
49	637	0.76	92.70
50	636	0.75	93.45
51	606	0.72	94.17
52	598	0.71	94.88
53	605	0.72	95.60
54	590	0.70	96.30
55	563	0.67	96.97
56	506	0.60	97.57
57	543	0.64	98.21
58	466	0.55	98.76
59	536	0.64	99.40
60	509	0.60	100.00

Tables 20 and 21 show the number of students who had multiple wrong-to-right changes in 10 seconds or less. If the frame of reference for considering this information is that a person, other than the student herself, changes the answer in quick succession, then this table helps reveal patterns of quick changes. Very small numbers of students more than five wrong-to-right changes in less than 10 seconds.

Table 20

Wrong-to-Right Changes Per Student Changed in 10 seconds or Less Elapsed Screen Time – English Language Arts

Wrong-to- Right Changes per Student	Count	Percent	Cumulative Percent
1	9,060	76.56	76.56
2	1,930	16.31	92.87
3	504	4.26	97.13
4	188	1.59	98.72
5	74	0.63	99.34
6	35	0.30	99.64
7	16	0.14	99.77
8	12	0.10	99.87
9	3	0.03	99.90
10	3	0.03	99.92
11	4	0.03	99.96
12	1	0.01	99.97
13	2	0.02	99.98
16	1	0.01	99.99
18	1	0.01	100.00

Table 21

Wrong-to-Right Items Per Student Changed in 10 seconds or Less Elapsed Screen Time – Mathematics

Wrong-to-Right Changes per Student	Count	Percent	Cumulative Percent
1	17,438	76.81	76.81
2	3,646	16.06	92.87
3	875	3.85	96.72
4	293	1.29	98.01
5	129	0.57	98.58
6	86	0.38	98.96
7	53	0.23	99.19
8	33	0.15	99.34
9	33	0.15	99.48
10	23	0.10	99.59
11	22	0.10	99.68
12	14	0.06	99.74
13	13	0.06	99.80
14	10	0.04	99.85
15	12	0.05	99.90
16	6	0.03	99.93
17	4	0.02	99.94
18	2	0.01	99.95
19	3	0.01	99.96
21	2	0.01	99.97
22	2	0.01	99.98
23	1	0.00	99.99
24	1	0.00	99.99
26	1	0.00	100.00
35	1	0.00	100.00

Modeling Answer-Changing Variables

Next, a series of models were fit to the answer-changing variables total answer changes, wrong-to-right answer changes, and the proportion of wrong-to-right to total answer changes. The models included the Poisson regression and negative binomial regression. Student demographic predictor variables included grade level, gender, comprehensive race, free or reduced lunch status, whether or not students received English as Second Language Services, and whether or not students received special education services.

As described in Chapter 2, Poisson models for count data can be hindered by overdispersion since the model assumptions state that the mean and variance must be equal across all points on the distribution. Poisson model overdispersion was calculated by dividing each model's residual deviance by its residual degrees of freedom (Mazerolle, 2004). Results indicated that this value for overdispersion was greater than one for each Poisson model, thus Poisson-based results were not further evaluated or interpreted.

As discussed in the Chapter Three, the Akaike information criterion (AIC), delta AIC, and Aikaike weights were used to compare the competing models. Results for the top three models in each model fitting activity are listed and described below. Results for English language arts and mathematics are described separately.

English Language Arts

In the first group of analyses for the English language arts assessment, the dependent variable Total Answer Changes was evaluated using negative binomial regression. Models were applied separately to the total population of students and then to a sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 22.

Table 22

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	483,804.71	17	483,840.71	0.00	0.91
grade, gender, race, lunch	3	483,814.54	15	483,846.54	5.82	0.05

grade, gender, race, lunch, ESOL	2	483,813.96	16	483,847.96	7.25	0.02
--	---	------------	----	------------	------	------

Results of Table 22 indicate that Model 1, the global model, is the best model among the pool of negative binomial models with an Akaike weight of 0.91. Model 1 is 18.2 times more likely than Model 3 (evidence ratio 0.91 / 0.05) and 45.5 times more likely than Model 2 (evidence ratio 0.91 / 0.02). Model 1 contains the variables grade, gender, race, lunch status, ESOL status, and special education status with no interaction effects represented.

In the next group of analyses, only students who had changed one or more answers were analyzed. Thus, students with zero answer changes were removed from the population of examinees. Results of this analysis are listed in Table 23.

Table 23

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	234,440.51	17	234,476.51	0.00	0.80
gender, race, lunch, ESOL	8	234,458.69	10	234,480.69	4.17	0.10
race, lunch, ESOL	11	234,460.77	9	234,480.77	4.26	0.09

Results of Table 23 show that Model 1, the global model, is also the best model in the pool of negative binomial models with an Akaike weight of 0.80. Model 1 is 8 times more likely than Model 8 (evidence ratio 0.80 / 0.10) and 8.88 times more likely than Model 11 (evidence

ratio 0.80 / 0.09). Model 1 contains the variables grade, gender, race, lunch status, ESOL status, and special education status with no interaction effects represented.

In the second group of analyses for the English language arts test data, the dependent variable wrong-to-right answer changes was evaluated using negative binomial regression. Models were applied separately to the total population of students and then to a sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 24.

Table 24

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
race, lunch, ESOL	11	330,422.31	9	330,442.31	0.00	0.53
gender, race, lunch, ESOL	8	330,421.12	10	330,443.12	0.81	0.36
grade, gender, race, lunch	3	330,415.33	15	330,447.33	5.01	0.04

Results of Table 24 show that Model 11 is also the best model in the pool of negative binomial regression models with an Akaike weight of 0.53. Model 11 is 1.47 times more likely than Model 8 (evidence ratio 0.53 / 0.36) and 13.25 times more likely than Model 3 (evidence ratio 0.53 / 0.04). Model 11 contains the variables race, lunch status, and ESOL status with no interaction effects represented.

Next, only students who had changed one or more answers were modeled. Thus, students with zero answer changes were removed from the population of examinees. Results of this analysis are listed in Table 25.

Table 25

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
race, lunch, ESOL	11	234,440.51	9	185,295.48	0.00	0.64
gender, race, lunch, ESOL	8	234,458.69	10	185,297.20	1.72	0.27
grade, race, lunch, ESOL	4	234,460.77	15	185,300.86	5.38	0.04

Results of Table 25 indicate that Model 11 is the best model in the pool of negative binomial models with an Akaike weight of 0.64. Model 11 is 2.37 times more likely than Model 8 (evidence ratio $0.64 / 0.27$) and 16 times more likely than Model 4 (evidence ratio $0.64 / 0.04$). Model 11 contains the variables race, lunch, and ESOL status with no interaction effects among the variables represented.

In the last group of analyses with the English language arts test data, the dependent variable proportion of wrong-to-right to total changes was evaluated using negative binomial regression with an offset variable consisting of the log of the total change count. For this set, models were applied only to the sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 26.

Table 26

Results of AIC analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	160,258.00	17	160,294.00	0.00	0.91

race, lunch, ESOL	11	160,280.02	9	160,300.02	6.02	0.05
gender, race, lunch, ESOL	8	160,279.71	10	160,301.71	7.71	0.02

Results of Table 26 indicate that Model 1, the global model, is clearly the best model among the group of binomial models with an Akaike weight of 0.91. Model 1 contains the variables grade, gender, race, lunch status, ESOL status, and special education status with no interaction effects represented.

Mathematics

In the first group of analyses for mathematics, the dependent variable Total Answer Changes was evaluated using negative binomial regression. Models were applied separately to the total population of students and then to a sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 27.

Table 27

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	809,662.86	3	809,670.86	0.00	0.23
grade, race, lunch, ESOL	4	809,638.95	15	809,670.95	0.09	0.22
grade, gender, race, lunch, ESOL	2	809,637.76	16	809,671.76	0.90	0.15

The results in Table 27 indicate that Model 9 and Model 4 are very close in terms of Akaike statistics. The difference in the AIC values is very small at 0.09. In fact, Model 9 is only 1.04 times more likely than Model 4 to be the best model (evidence ratio $0.23 / 0.22$). Relative to

Model 2, Model 9 is only 1.53 times more likely to be the best model (evidence ratio 0.23 / 0.15).

Model 9 is only 1.77 times more likely than the global model, Model 1, to be the best model (evidence ratio 0.23 / 0.13).

In the next group of analyses, only students who had changed one or more answers were analyzed. Thus, students with zero answer changes were removed from the population of examinees. Results of this analysis are listed in Table 28.

Table 28

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	541,728.15	3	541,736.15	0.00	0.89
race, lunch, ESOL	11	541,722.52	9	541,742.52	6.37	0.04
grade, race, lunch, ESOL	4	541,711.61	15	541,743.61	7.46	0.02

The results of Table 28 indicate that Model 9 is the best model in the pool with an Akaike weight of 0.89. Model 9 is 22.25 times more likely than Model 11 (evidence ratio 0.89 / 0.04). Model 9 contains only the variables lunch and ESOL status, with no interaction effects included in the model.

In the second group of analyses, the dependent variable wrong-to-right answer changes was evaluated using negative binomial regression. Models were applied separately to the total population of students and then to a sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 29.

Table 29

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	624,650.57	3	624,658.57	0.00	0.29
grade, gender, race, lunch	3	624,626.73	15	624,658.73	0.16	0.26
grade, gender, race, lunch, ESOL	2	624,625.61	16	624,659.61	1.04	0.17

The results of Table 29 indicate that Model 9 is the best model among the pool of negative binomial models with an Akaike weight of 0.29. However, Models 3 and 2 are quite close, with Akaike weights of 0.26 and 0.17 respectively. Model 9 is 1.12 times more likely than Model 3 (evidence ratio $0.29 / 0.26$), 1.71 times more likely than Model 2 (evidence ratio $0.29 / 0.17$), and 2.07 times more likely than Model 4 (evidence ratio $0.29 / 0.14$). Model 9 has fewer variables than Models 3, 2, or 4 with only lunch and ESOL status present and no interaction effects represented.

In the next group of analyses, only students who had changed one or more answers were analyzed. Thus, students with zero answer changes were removed from the population of examinees. Results of this analysis are listed in Table 30.

Table 30

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	541,728.15	3	447,450.50	0.00	0.91
grade, gender, race, lunch	3	541,722.52	15	447,457.12	6.62	0.03

grade, race, lunch, ESOL	4	541,711.61	15	447,457.91	7.41	0.02
-----------------------------	---	------------	----	------------	------	------

Results of Table 30 also indicate that Model 9 is the best model of the pool of negative binomial regression models, with an Akaike weight of 0.91. Model 9 is 30.33 times more likely than Model 3 (evidence ratio $0.91 / 0.03$) and 45.5 times more likely than Model 4 (evidence ratio $0.91 / 0.02$).

In the last group of analyses with the mathematics test data, the dependent variable proportion of wrong-to-right to total changes was evaluated using negative binomial regression with an offset variable consisting of the log of total item change counts. For this set, models were applied only to the sub-population of students who made at least one change to any item on the test. Results of this analysis are listed in Table 31.

Table 31

Results of AIC Analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	356,788.13	3	356,788.13	0.00	0.55
ESOL	10	356,788.88	2	356,788.88	0.75	0.38
race, lunch, ESOL	11	356,792.95	9	300,868.78	4.81	0.05

The results of Table 31 indicate that Model 9 is the best model among this pool of negative binomial regression models with an Akaike weight of 0.55. Model 9 is 1.45 times more likely than Model 10 (evidence ratio $0.55 / 0.38$), and 11 times more likely than Model 11 (evidence ratio $0.55 / 0.05$). Model 9 contains the variables lunch status and ESOL status with no interaction effects represented.

Item-Level Analysis

Next, an exploration of the data with the item as the unit of analysis was completed. Across all grades, the data included 959 English language arts items and 549 mathematics items. The distribution of items across grades is listed in Table 32 for English language arts and Table 33 for mathematics. In this particular testing year, English language arts was not tested at the 10th grade level and mathematics was not tested at the 11th grade level.

Table 32

Number of Items per Grade, English Language Arts

	Grade-level							
	3	4	5	6	7	8	10	11
Number of Items	108	136	136	134	146	138	0	161

Table 33

Number of Items per Grade, Mathematics

	Grade-level							
	3	4	5	6	7	8	10	11
Number of Items	60	62	65	72	70	71	149	0

The number of students answering each English language arts item ranged from 985 to 28,995, with a mean of 7,399 (*SD* 4,608.72). The number of students answering each mathematics item ranged from 7,354 to 42,067, with a mean of 31,366 (*SD* 9,965.21).

Wrong-to-Right Changes per Item

Next, the wrong-to-right changes per item were calculated. For English language arts, the total number of changes per item ranged from 1 to 732 with an average of 71.41 (*SD* 62.93). For mathematics, the number of wrong-to-right changes per item ranged from 22 to 1,622, with a mean of 364.2 (*SD* 226.15). Figure 3 displays the distribution of wrong-to-right changes per item for English language arts and mathematics.

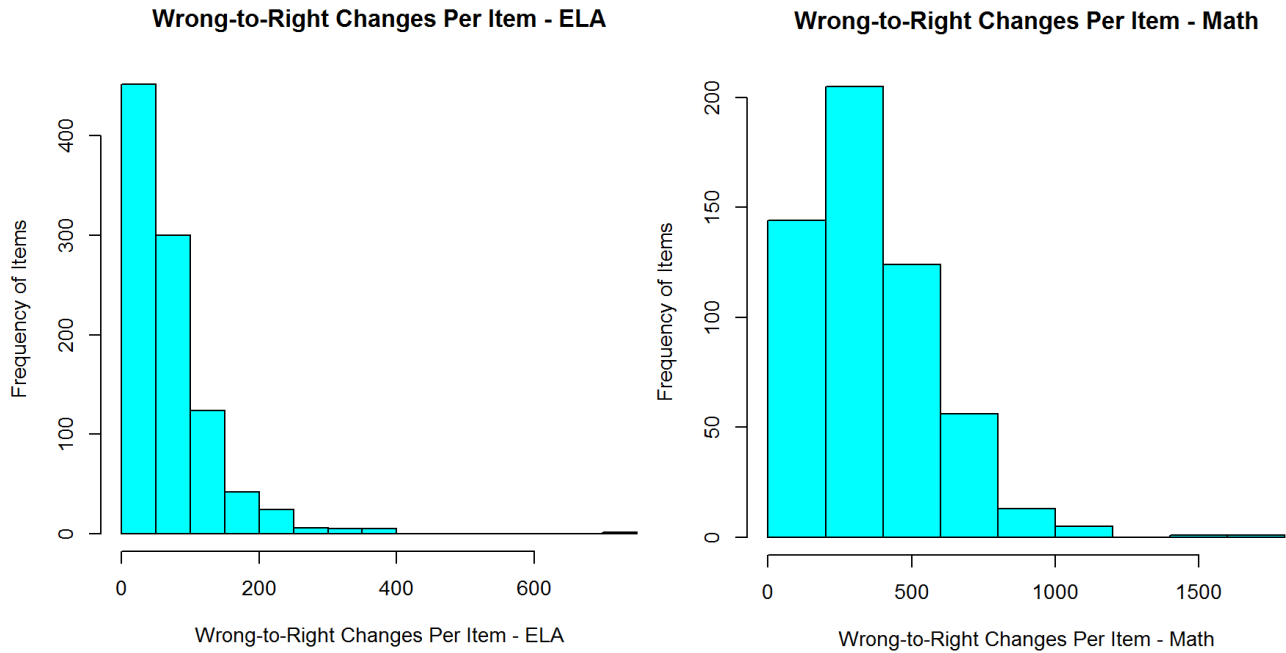


Figure 3. Frequency of wrong-to-right changes per item, English language arts and mathematics

Right-to-Wrong Changes per Item

As a comparison to the above wrong-to-right change information, the distribution of right-to-wrong changes per item (combined English language arts and mathematics) is displayed in Figure 4. The number of right-to-wrong changes per item ranged from 0 to 1,167, with a mean of 56.04 (SD 73.84). The distribution of right-to-wrong was also skewed right and reflects the fact that students change answers right to wrong less frequently than wrong to right.

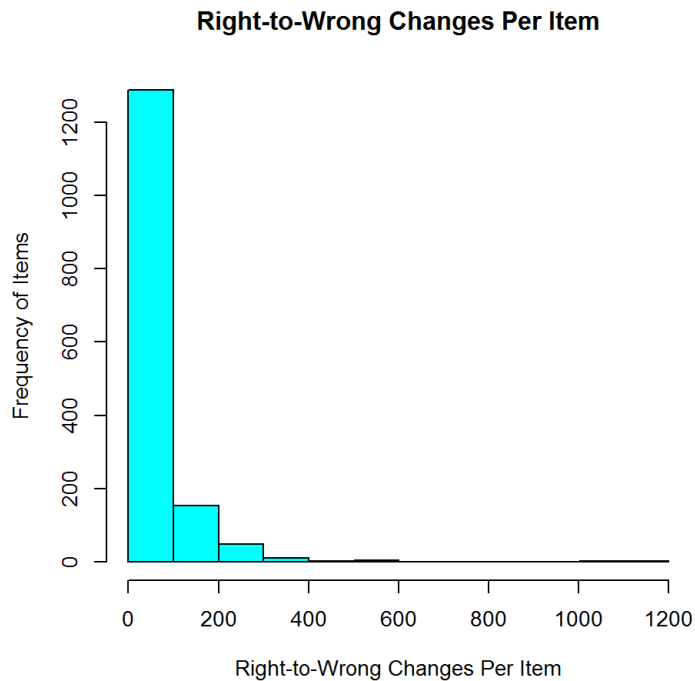


Figure 4. Frequency of right-to-wrong changes per item

Relationships Between Answer Changes and Item Difficulty

Given that answer-changing behavior may vary based on the difficulty of the item, the frequency of p-values across English language arts and mathematics items were plotted. In English language arts, the mean p-value was 0.80 (*SD* 0.13). For mathematics, the mean p-value was 0.74 (*SD* 0.16). Thus, both exams were relatively “easy” with both distributions skewed left. The distribution of p-values for English language arts and mathematics assessments are displayed in Figure 5.

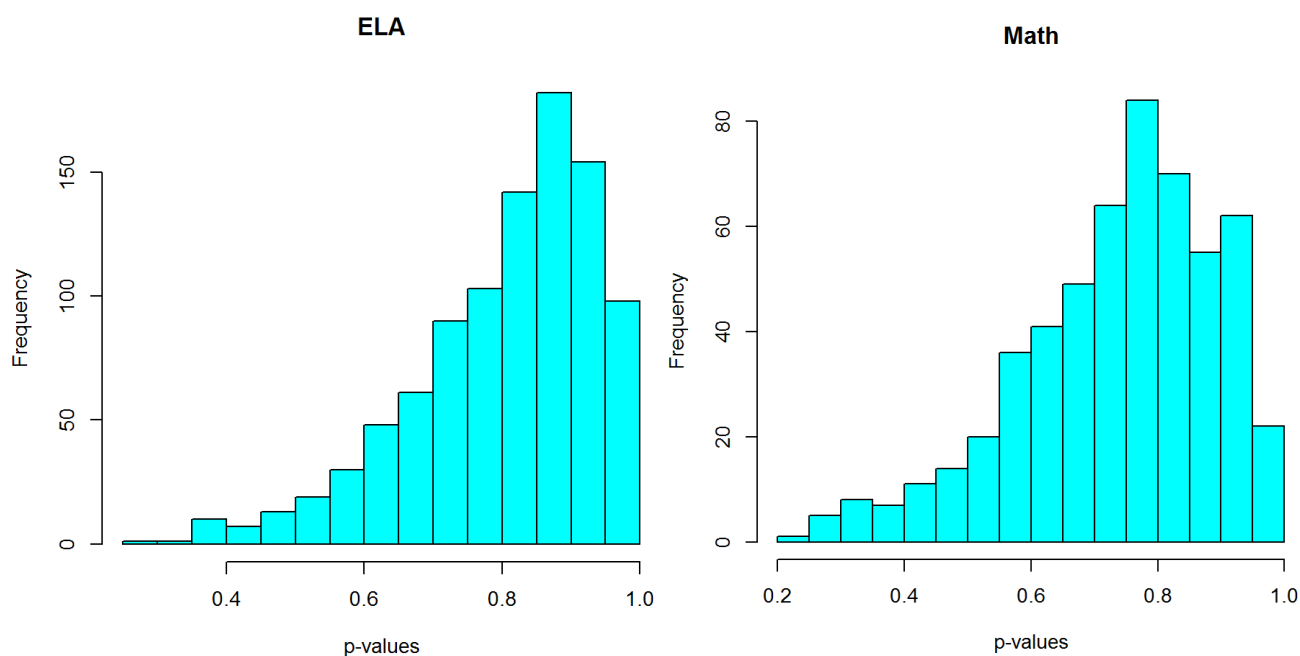


Figure 5. Distribution of p-values – English language arts and mathematics

In order to facilitate comparisons among items with varying numbers of students answering as well as to examine all items on the same scale, normalized p-values were calculated. Normalizing the p-values converts each p-value to a z-score on the standard normal distribution, with higher values indicating easier items.

Figure 6 below plots the relationship between normalized item p-values and total changes per item. For both English language arts and mathematics, it is clear that there is a negative relationship between p-value and total changes; more difficult items were changes more frequently than easier items.

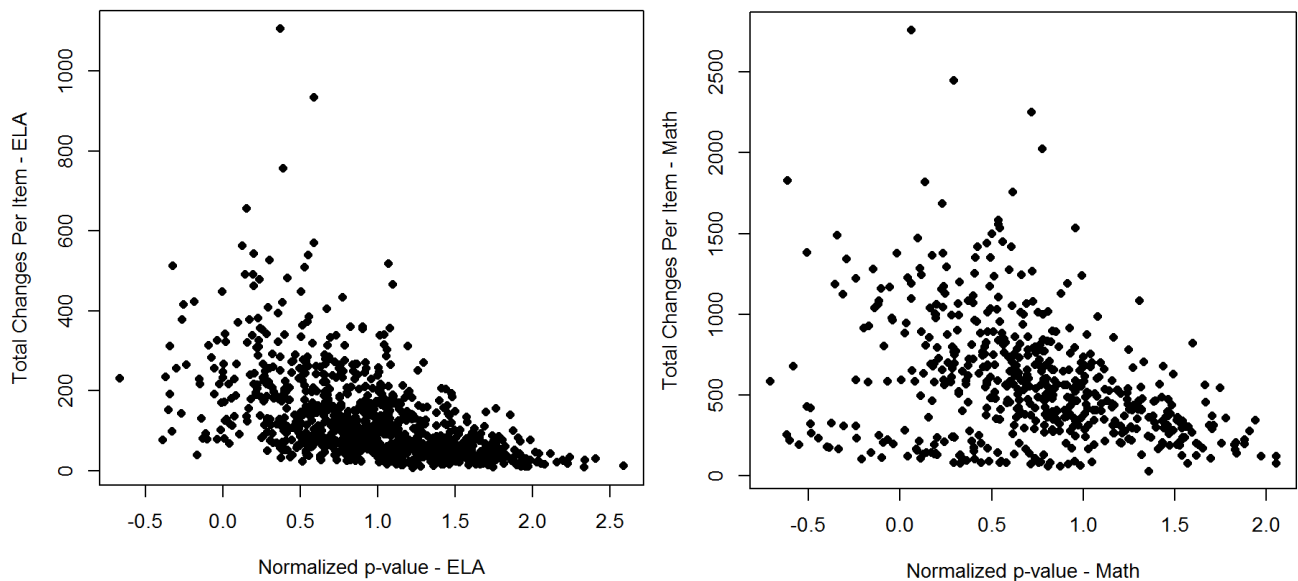


Figure 6. Plot of normalized of p-values versus total changes per item– English language arts and mathematics

Figure 7 below plots the relationship between normalized p-values and the proportion of wrong-to-right to total changes. The plot indicates that there is a positive relationship between easier items and higher proportions of wrong-to-right to total changes. Since correct items have higher p-values, and wrong-to-right items are by definition correct items, this positive relationship is an expected result.

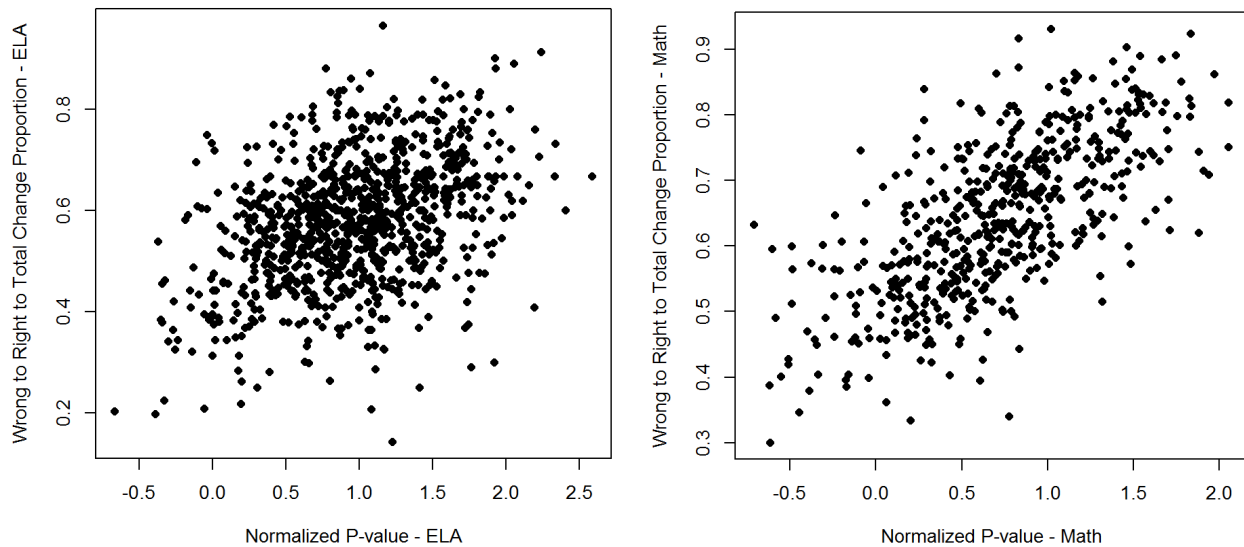


Figure 7. Plot of normalized p-values versus the proportion of wrong-to-right to total changes – English language arts and mathematics

Summary

This chapter presented the results of the data analysis procedures used to explore the study's research questions. Answer-changing frequencies were presented for total item changes, wrong-to-right changes, right-to-wrong changes. Response time distributions were also presented. Additionally, contingency tables allowed for comparison of different count variables. Various regression models were also applied, however dispersion in models prevented use of Poisson models. Results of negative binomial regression models were presented along with plots of the relationships between item difficulty and answer changing.

CHAPTER 5 - DISCUSSION

The purpose of this study was to explore and document answer-changing patterns of students grades 3-12 on computer-based English language arts and mathematics mandated state achievement tests. This chapter presents the results of the study as well as limitations and suggested areas of further study.

Answer-Changing Frequencies

Results of answer-changing frequency analysis showed that about 30% of English language arts examinees and about 57% of mathematics examinees changed at least one item. Qualls (2001) reported that about 50% of students across grades and content areas changed at least one item. Thus in this study the values for English language arts were slightly lower than that reported by Qualls, though the values for mathematics were more inline with Qualls' results. Additionally, the percentages in this study were overall lower than those reported by Benjamin et al. (1984), McMorris et al. (1991), and Geiger (1991) who reported values of 57-95%, 75%, and 97% respectively.

The average number of total answers changed in this study was 0.52 (*SD* 1.13) for English language arts and 1.31 (*SD* 2.04) for mathematics. These values are far lower than those reported by Jackson (1978) and McMorris and Weideman (1986). Additionally, Geiger reported large standard deviations, with some larger than means. Large standard deviations also occurred in this study, where standard deviations for English language arts and mathematics were both larger than their respective means. However, the means were consistent with results of Primoli, Liassou, Bishop, and Nhouyvanisvong (2011), who noted that answer changing overall was a relatively rare occurrence. This was easily inferred as well from the percentages of students changing four answers or less in this study. Out of all of the items taken, 1% of English language

arts examinees changed five or more items and 5% of mathematics examinees changed five or more items.

Results of this study also indicate that of the items that were changed, wrong-to-right changes were also rare occurrences, especially in English language arts where only 20% of students had one or more wrong-to-right change. Additionally, the means for wrong-to-right changes were lower than that of total changes, but also with wide variances. The mean for English language arts wrong-to-right changes was 0.27 (*SD* 0.67) and mathematics was 0.78 (*SD* 0.78).

Patterns Of Wrong-To-Right Changes

In the context of exploring potential cheating behaviors among answer changers, several views of the wrong-to-right answer changing patterns may be helpful for determining which students might require further examination. For example, while one could apply a straight flagging rule to basic wrong-to-right counts, a contingency table (such as Table 12) which plots final wrong-to-right changes against total item changes provides a little more context. One can see across the diagonals when a student has changed a 1:1 proportion of wrong-to-right changes to total changes. Additionally from this view, one can see just how infrequently high numbers of changes to total changes actually occurs. The right-to-wrong changes plotted against the wrong-to-right changes also presents an interesting view. As mentioned, one might be less concerned with a student who had high-numbers of both types, but more concerned about high-numbers of wrong-to-right and low right-to-wrong.

Response Time

Time adds another dimension to exploration of student answer changes. If one wishes to consider the possibility of a third party changing a student answers, one may think that the

changes could be made in quick succession. In this population of examinees, documentation of response times showed that about half of all wrong-to-right changes occurred in five seconds of screen time or less for English language arts, and about 14 seconds or less for mathematics. Note that English language arts had lower overall answer changes than mathematics. When looking at quick response times for multiple items for a single student, very few students had more than five wrong-to-right changes in less than ten seconds. It is difficult to infer from response time alone if aberrant behavior is present in a student's item responses. However, if one identifies a high wrong-to-right to total change proportion or a low right-to-wrong to wrong-to-right proportion, adding an analysis of the response times could become an additional step in exploration of a potential problem.

Item-Level Answer-Changing Frequencies

An item-level exploration of the data was also completed. Results showed that the test was overall all relatively easy, with mean p-values of 0.80 (*SD* 0.13) and 0.74 (*SD* 0.16) for English language arts and mathematics respectively. Additionally, for both English language arts and mathematics, there was a negative relationship between normalized p-values and total answer changes. These results were consistent several other, older studies that also reported easier items were changed less frequently (Benjamin et al, 1984; Jacobs 1972; Green, 1981; McMorris & Weideman, 1986).

Related to time at the item-level, average screen-times for each item in general, average screen time per wrong-to-right change, and average screen time per right-to-wrong change were consistent within each content area. On average, students take about 29 seconds to answer an item or change an item in English language arts, and they take on average about 46 seconds to answer or change an answer in math. Because of the wide-variability though in these averages,

as well as the right skew of the response-time distributions, it is difficult to tell from this study if screen-time on its own is a useful indicator of potential cheating behavior. Again other variables could be more highly influencing the response time and could be evaluated more fully.

Modeling Results

In this study, Poisson regression and negative binomial regression models were fit to the independent variables of total answer changes, wrong-to-right answer changes, and the proportion of wrong-to-right to total answer changes. Additionally, various models including several demographic variables were compared using the information-theoretic approach. Results indicated that overdispersion in the models precluded the use of Poisson regression in both English language arts and mathematics. These results are consistent with that of Bishop et al. (2011), who also found that use of the negative binomial model was a better fit to answer-changing data than the Poisson model. Note that Bishop et al. used data from paper-pencil assessments, rather than computer-based assessments, as was the case with this study.

The demographic variables themselves were not useful in explaining answer-changing behaviors, for any of the independent variables examined. While the AIC values helped to determine which models were the leading models, especially considering the large and thus possibly over-powered sample size, the estimated coefficients were too similar to be practically useful. Thus, the variables do not appear to be associated with this data, however other variables could be and should be explored more fully. The results of this analysis do however, support the notion that flagging rules that depend on the assumptions of a normal distribution could lead to spurious results.

Additional Limitations

Another component of this exploration was response time for items and answer-changes. Response time was indirectly measured by the amount of time a student was on the screen before surfing away. A much more direct measure of response-time could be derived from actual time stamps as answers are written to server database tables. This direct measure would allow for more precision in the response-time values. Database time stamps were not available for this study.

This study helps to build upon prior work and creates more knowledge about patterns of student-answer changing. It does not provide evidence that any aberrant behavior actually occurred. High-levels of answer changing could be explained by other factors that were not a focus of this exploration. Student could change answers to pass time, to appear engaged in the test, or as a legitimate test-taking strategy.

On a broader level, it is conceivable that regular and widely publicized answer-changing analysis could potentially become a deterrent for third-party manipulation of student answers. A negative by-product of such a deterrent, could be educators telling students to stick with their first answer so as to not raise suspicion. Such a guideline would contradict 70 years of research which has consistently found that changing answers generally works for a student, not against.

Future Research

While documenting student demographic characteristics as potential sources of variability in the answer-changing counts is helpful, it may not fully tell the story. Future research could help to surface other sources of variability. For example, as a post-hoc study, students who had answer-changing behaviors higher-than the mean could be interviewed and asked why particular answers resulted in changes. Additionally, further work could explore more

specific item characteristics relative to answer changing such as item content, item phrasing, or text complexity.

A significant portion of this work examined answer changing at the student level. Because students generally take tests as classes, the data are nested at the classroom and building levels. Due to limitations in the student information that was available in this data set, examination by classroom level was not possible. Future research should evaluate classroom-level and building-level behaviors by applying hierarchical modeling approaches. Such an examination could provide more insight into potential tampering with student responses by third-parties.

Additionally, this study did not examine answer-changing as a function of the position student responses, e.g. were answers changed at the beginning of the student's linear path through the test or at the end of the student's path through the test. Future research could document student answer patterns more fully as a potential step to include in addition to evaluation of counts and response times.

Conclusion

This study sought to explore student answer-changing in order to document the construct more fully. Results are consistent with prior research that has indicated that large numbers of answer changes are rare occurrences that could warrant further exploration. Additionally, results can be treated as a baseline for comparison of future testing windows.

These results might be useful to state education agencies (SEAs) as they consider policies regarding the triggering of investigations related to score-changing behaviors. Much of this type of analysis should be approachable by regular SEA staff with access to count data and spreadsheet software. Additionally, the modeling results of this study indicate that state

education agencies should use caution if working with vendors that use such flagging rules as the only approach to evaluating student response data. SEAs should ask test-security vendors to provide evidence that chosen methods are a good fit with the data and whether or not analytical assumptions are met.

In sum, answer-changing continues to be a valuable tool as part of a broader test-integrity approach. Future research could document the construct more fully to bring more understanding of student-answer changing behavior.

REFERENCES

- Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for gulf Arab students. *Language Testing*, 22(4), 509-531. doi: 10.1191/0265532205lt317oa
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives*, 1-36.
- Beck, M. D. (1978). The effect of item response changes on scores on an elementary reading achievement test. *The Journal of Educational Research*, 71(3), 153-156. doi: 10.1080/00220671.1978.10885059
- Bello, M., & Toppo, G. (2011, September 13). Few states examine test erasures, *USA Today*. Retrieved from <http://usatoday30.usatoday.com/news/education/story/2011-09-12/states-analyze-test-erasures/50376902/1>
- Benjamin, L. T., Jr., Cavell, T. A., & Shallenberger, W. R., III,. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, 11(3), 133-141.
- Best, J. B. (1979). Item difficulty and answer changing. *Teaching of Psychology*, 6(4), 228-230. doi:10.1207/s15328023top0604_10
- Bishop, N. S., Liassou, D., Bulut, O., Dong, G. S., & Stearns, M. (2011, April). *Paper Four: Application and comparison of alternative procedures*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

- Bishop, N. S., Liassou, D., Bulut, O., Seo, D. G., & Bishop, K. (2011, April). *Paper Three: Modeling erasure behavior*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). New York: Springer.
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data* (Vol. 30): Cambridge University Press.
- Casteel, C. A. (1991). Answer changing on multiple-choice test items among eighth-grade readers. *Journal of Experimental Education*, 59(4), 300-309.
- Cizek, G. J. (1999). *Cheating on tests : How to do it, detect it, and prevent it*. Mahwah, NJ: L. Erlbaum Associates.
- Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2), 121-136. doi: 10.1080/00223890802634175
- Crocker, L., & Benson, J. (1977). Effects of examinee response changes on item and test characteristics. Gainesville, FL: Florida University-Institute for Development of Human Resources.
- Duncan, A. (2011). Key policy letters from the education secretary or deputy secretary Retrieved November 5, 2011, from <http://www2.ed.gov/policy/elsec/guid/secletter/110624.html>
- Erasure Analysis, La. Admin. Code Tit. 28 Part CXI Bulletin 118-Statewide Assessment Standards and Practices § 309 (2012).

- Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3), 392-404. doi: 10.1037/0033-2909.118.3.392
- Geiger, M. A. (1991). Changing multiple-choice answers: Do students accurately perceive their performance? *Journal of Experimental Education*, 59(3), 250-257. doi: 10.1080/00220973.1991.10806564
- Geiger, M. A. (1997). An examination of the relationship between answer changing, testwiseness, and examination performance. *Journal of Experimental Education*, 66(1), 49-60. doi: 10.1080/00220979709601394
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research*. Chicago,: Aldine Pub. Co.
- Green, K. (1981). Item-response changes on multiple-choice tests as a function of test anxiety. *Journal of Experimental Education*, 49(4), 225-228.
- Halkitis, P. N. (1996, April). *Estimating testing time: The effects of item characteristics on response latency*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.
- Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91-114). Mahwah, N.J.: L. Erlbaum.
- Impara, J. C., Kingsbury, G., Mayes, D., & Fitzgerald, C. (2005, April). *Detecting cheating in computer adaptive tests using data forensics*. Paper presented at the Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.

- Jackson, P. F. (1978). Answer changing on objective tests. *Journal of Educational Research*, 71(6), 313-315. doi: 10.1080/00220671.1978.10885097
- Jacobs, S. S. (1972). Answer changing on objective tests: Some implications for test validity. *Educational and Psychological Measurement*, 32(4), 1039-1044. doi: 10.1177/001316447203200420
- Kato, K., & Bart, W. M. (2010). Poisson distribution. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 1042-1045). Los Angeles, CA: SAGE.
- Kimmel, E. W. (1997, June). *Unintended consequences or testing the integrity of teachers and students*. Paper presented at the Annual Assessment Conference of the Council of Chief State School Officers, Colorado Springs, CO.
- Mathews, C. O. (1929). Erroneous first impressions on objective tests. *Journal of Educational Measurement*, 20(4), 280-286. doi: 10.1037/h0071721
- Matter, M. K. (1986, April). *Eenie, meenie, minie, mo--change this answer--yes or no?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Mazerolle, M. J. (2004). *Appendix 1: Making sense out of Akaike's Information Criterion (AIC): Its use and interpretation in model selection and inference from ecological data*. (Doctoral thesis), Université Laval. Retrieved from <http://avesbiodiv.mnecn.csic.es/estadistica/senseaic.pdf>
- McMorris, R. F., & Leonard, G. (1976). Item response changes and cognitive style. Albany, NY: SUNY.
- McMorris, R. F., Schwarz, S. P., Richlichi, R. V., Fischer, M., Buczek, N. M., Chevalier, C. L., & Meland, K. A. (1991). *Why do young students change answers on tests?* Paper

- presented at the 22nd annual Northeastern Educational Research Association Conference, Ellenville, NY (p. 22). (ERIC – ED342803)
- McMorris, R. F., & Weideman, A. H. (1986). Answer changing after instruction on answer changing. *Measurement and Evaluation in Counseling and Development*, 19(2), 93-101.
- Milliken, P. J. (2010). Grounded theory. In N. Salkind (Ed.), *Encyclopedia of research design* (Vol. 1, pp. 548-553). Thousand Oaks, CA: Sage Publications, Inc.
- Mueller, D. J., & Shwedel, A. (1975). Some correlates of net gain resultant from answer changing on objective achievement test items. *Journal of Educational Measurement*, 12(4), 251-254. doi: 10.1111/j.1745-3984.1975.tb01026.x
- Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement*, 14(1), 9-14. doi:10.1111/j.1745-3984.1977.tb00023.x
- National Center for Education Statistics. (2013). Testing integrity symposium: Issues and recommendations for best practice. Washington DC: United States Department of Education. Retrieved from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2013454>.
- National Council on Measurement in Education. (2012). Testing and data integrity in the administration of statewide student assessment programs. Madison, WI: Retrieved from <http://ncme.org/ncme-news/testing-and-data-integrity/>.
- Nickerson, R. S. (2010). Inference: Deductive and inductive. In N. Salkind (Ed.), *Encyclopedia of research design* (Vol. 2, pp. 593-596). Thousand Oaks, CA: Sage Publications, Inc.
- Nieswiadomy, R. M., Arnold, W. K., & Garza, C. (2001). Changing answers on multiple-choice examinations taken by baccalaureate nursing students. *Journal of Nursing Education*, 40(3), 142-143.

- Otterman, S. (2011, September 23). State says it analyzed test erasures for cheating; 62 schools proved suspect, *The New York Times*. Retrieved from <http://www.nytimes.com/2011/09/24/nyregion/in-reversal-new-york-state-says-it-used-erasure-analysis-to-detect-cheating.html?pagewanted=all>
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011, April). *Paper Two: Erasure descriptive statistics and covariates*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Prinsell, C. P., Ramsey, P. H., & Ramsey, P. P. (1994). Score gains, attitudes, and behavior changes due to answer-changing instruction. *Journal of Educational Measurement*, 31(4), 327-337. doi: 10.1111/j.1745-3984.1994.tb00450.x
- Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, 20(1), 9-16.
doi:10.1111/j.1745-3992.2001.tb00053.x
- Reile, P. J., & Briggs, L. J. (1952). Should students change their initial answers on objective-type tests? More evidence regarding an old problem. *Journal of Educational Psychology*, 110-115. doi: 10.1037/h0057463
- Rideau, S. (2009). *Teachers cheating on standardized achievement tests: Perceived causes and effects*. Doctor of Education Dissertation, Arizona State University. Dissertation Abstracts database. (3357277)

- Schaeffer, G. A. (1995). The introduction and comparability of the computer adaptive GRE general test. (GRE Board Professional Report No. 88-08aP). Princeton, NJ: Educational Testing Service.
- Schiliro, K. (2010, February 18). Questionable CRCT answer sheets at MCES?: “Minimal concern” category by two-tenths of a percent, *Morgan County Citizen*. Retrieved from <http://www.morgancountycitizen.com/?q=node/12946>
- Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3), 213-232. doi: 10.1111/j.1745-3984.1997.tb00516.x
- Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks* (Vol. 97). Newtown, PA: Law School Admission Council.
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. Mills, M. Potenza, J. Fremer & W. Ward (Eds.), *Computer-based testing : Building the foundation for future assessments* (pp. 237–266). Mahwah, N.J.: L. Erlbaum Associates.
- Segrin, C. (2010). Multiple regression. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 844-849). Los Angeles, CA: Sage.

- Severson, K. (2011, July 5). Systematic cheating is found in Atlanta's school system, *The New York Times*. Retrieved from <http://www.nytimes.com/2011/07/06/education/06atlanta.html>
- Skinner, N. F. (1983). Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology*, 10(4), 220-222. doi: 10.1207/s15328023top1004_9
- Skorupski, W. P., & Wainer, H. (2013). *The "P" you really want to know: Why you should detect cheating the Bayesian way*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Smith, R. W. (2000). *An exploratory analysis of item parameters and characteristics that influence item level response time*. Doctor of Philosophy Dissertation, University of Nebraska. Dissertation Abstracts database. (9973602)
- Stebbins, R. A. (2001). *Exploratory research in the social sciences*. Thousand Oaks, California.: Sage Publications.
- Swanson, D., Featherman, C., Case, S., Luecht, R., & Nungester, R. (1999). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.
- Tang, W., He, H., & Tu, X. M. (2012). *Applied categorical and count data analysis*. Boca Raton: CRC Press.

- Tiemann, G. C., & Kingston, N. M. (2012, May). *An exploration of answer-changing on a computer-based high-stakes achievement test*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Tiemann, G. C., & Kingston, N. M. (2013, October). *Answer changing and response time on computer-based, high-stakes achievement tests*. Poster presented at the 2nd Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI.
- Vogt, W. P. (1999). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.
- Wainer, H. (2012, May). *How to detect cheating badly*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Wise, S. L., & Kingsbury, G. (2006, April). *An investigation of item response time distributions as indicators of compromised NCLEX item pools*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

APPENDIX

Coefficients of selected models

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

English Language Arts

Table A.1

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	483,804.71	17	483,840.71	0.00	0.91

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.636724	0.011937	-53.339	< 2e-16 ***
GRADE4	-0.002249	0.014771	-0.152	0.87897
GRADE5	-0.022771	0.014772	-1.541	0.12320
GRADE6	-0.025302	0.014807	-1.709	0.08749 .
GRADE7	-0.042858	0.014896	-2.877	0.00401 **
GRADE8	-0.038865	0.015080	-2.577	0.00996 **
GRADE11	-0.047324	0.014538	-3.255	0.00113 **
GENDER1	-0.012410	0.007974	-1.556	0.11962
RACE20	0.019132	0.037963	0.504	0.61428
RACE21	0.018838	0.025379	0.742	0.45792
RACE22	0.027215	0.016597	1.640	0.10105
RACE23	-0.013907	0.013611	-1.022	0.30690
RACE25	0.009513	0.019773	0.481	0.63042
RACE26	-0.029151	0.102365	-0.285	0.77582
LUNCH31	0.016130	0.008768	1.840	0.06583 .
ESOL31	0.011986	0.017594	0.681	0.49571
SPED31	-0.043256	0.014220	-3.042	0.00235 **

Table A.2

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	234,440.51	17	234,476.51	0.00	0.80

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.5352149	0.0086348	61.983	< 2e-16 ***
GRADE4	0.0102894	0.0106803	0.963	0.335343
GRADE5	0.0008027	0.0107105	0.075	0.940255
GRADE6	0.0151794	0.0107404	1.413	0.157566
GRADE7	-0.0007700	0.0108265	-0.071	0.943301
GRADE8	0.0071179	0.0109597	0.649	0.516037
GRADE11	-0.0021383	0.0105702	-0.202	0.839688
GENDER1	-0.0064473	0.0058022	-1.111	0.266490
RACE20	0.0210615	0.0275308	0.765	0.444262
RACE21	0.0075554	0.0183652	0.411	0.680780
RACE22	0.0217268	0.0120073	1.809	0.070379 .
RACE23	-0.0121553	0.0098922	-1.229	0.219156
RACE25	0.0112168	0.0143586	0.781	0.434688
RACE26	-0.0509768	0.0748244	-0.681	0.495691
LUNCH31	0.0107083	0.0063851	1.677	0.093529 .
ESOL31	0.0037591	0.0127283	0.295	0.767740
SPED31	-0.0383383	0.0104303	-3.676	0.000237 ***

Table A.3

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, English Language Arts

Model	Model ID	Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
race, lunch, ESOL	11	330,422.31	9	330,442.31	0.00	0.53
Coefficients:						
	Estimate	Std. Error	z value	Pr(> z)		
(Intercept)	-1.304255	0.006638	-196.480	<2e-16 ***		
RACE20	0.018216	0.044996	0.405	0.686		
RACE21	0.008283	0.030240	0.274	0.784		
RACE22	-0.014354	0.019901	-0.721	0.471		
RACE23	-0.019346	0.016194	-1.195	0.232		
RACE25	0.002376	0.023511	0.101	0.920		
RACE26	0.037928	0.119384	0.318	0.751		
LUNCH31	0.007050	0.010352	0.681	0.496		
ESOL31	-0.015072	0.021014	-0.717	0.473		

Table A.4

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
race, lunch, ESOL	11	234,440.51	9	185,295.48	0.00	0.64
Coefficients:						
	Estimate	Std. Error	z value	Pr(> z)		
(Intercept)	-0.0982741	0.0054141	-18.152	<2e-16 ***		
RACE20	0.0228681	0.0365726	0.625	0.532		
RACE21	-0.0028485	0.0246076	-0.116	0.908		
RACE22	-0.0205014	0.0162543	-1.261	0.207		
RACE23	-0.0169753	0.0131929	-1.287	0.198		
RACE25	0.0031532	0.0191557	0.165	0.869		
RACE26	0.0144000	0.0967181	0.149	0.882		
LUNCH31	0.0005719	0.0084496	0.068	0.946		
ESOL31	-0.0270270	0.0171198	-1.579	0.114		

Table A.5

Results of AIC analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, English Language Arts

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
grade, gender, race, lunch, ESOL, sped	1	160,258.00	17	160,294.00	0.00	0.91

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.651561	0.011499	-56.665	< 2e-16 ***
GRADE4	0.003781	0.014233	0.266	0.790514
GRADE5	0.029842	0.014177	2.105	0.035298 *
GRADE6	0.003818	0.014307	0.267	0.789551
GRADE7	0.017642	0.014372	1.228	0.219620
GRADE8	0.022051	0.014524	1.518	0.128942
GRADE11	0.027763	0.013994	1.984	0.047264 *
GENDER1	0.001947	0.007679	0.253	0.799910
RACE20	0.000666	0.036288	0.018	0.985357
RACE21	-0.010189	0.024428	-0.417	0.676612
RACE22	-0.041146	0.016148	-2.548	0.010835 *
RACE23	-0.004668	0.013132	-0.355	0.722236
RACE25	-0.007744	0.019009	-0.407	0.683726
RACE26	0.067217	0.095992	0.700	0.483781
LUNCH31	-0.010118	0.008462	-1.196	0.231811
ESOL31	-0.029019	0.017023	-1.705	0.088260 .
SPED31	0.047760	0.013557	3.523	0.000427 ***

Mathematics

Table A.6

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, All Students, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	809,662.86	3	809,670.86	0.00	0.23

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.282627	0.003663	77.159	< 2e-16	***
LUNCH31	-0.013203	0.005701	-2.316	0.020561	*
ESOL31	-0.036036	0.010126	-3.559	0.000373	***

Table A.7

Results of AIC Analysis for Competing Negative Binomial Regression Models – Total Answer Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	541,728.15	3	541,736.15	0.00	0.89

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.849372	0.002823	300.895	< 2e-16	***
LUNCH31	-0.013955	0.004400	-3.172	0.00152	**
ESOL31	-0.023178	0.007863	-2.948	0.00320	**

Table A.8

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, All Students, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	624,650.57	3	624,658.57	0.00	0.29

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.235540	0.004127	-57.073	< 2e-16 ***
LUNCH31	-0.020967	0.006434	-3.259	0.00112 **
ESOL31	-0.026025	0.011435	-2.276	0.02285 *

Table A.9

Results of AIC Analysis for Competing Negative Binomial Regression Models – Wrong-to-Right Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	541,728.15	3	447,450.50	0.00	0.91

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.331208	0.003476	95.272	< 2e-16 ***
LUNCH31	-0.021726	0.005429	-4.002	6.28e-05 ***
ESOL31	-0.013160	0.009689	-1.358	0.174

Table A.10

Results of AIC Analysis for Competing Negative Binomial Regression Models – Proportion of Wrong-to-Right Changes to Total Changes, Only Students with Answer Changes, Mathematics

Model	Model ID	-2Log-likelihood	Number of Parameters (K)	AIC	Delta AIC (Δ_i)	Akaike weight (w_i)
lunch, ESOL	9	356,788.13	3	356,788.13	0.00	0.55

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.518161	0.002997	-172.899	<2e-16 ***
LUNCH31	-0.007771	0.004687	-1.658	0.0973 .
ESOL31	0.010001	0.008384	1.193	0.2329